NEURAL MECHANISMS OF PRONOUN RESOLUTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jixing Li

May 2019

NEURAL MECHANISMS OF PRONOUN RESOLUTION

Jixing Li, Ph.D.

Cornell University 2019

To understand how pronoun resolution is implemented in the brain, one first step is to describe the algorithm that performs the task. This thesis evaluated three computational models for pronoun resolution against brain activity time-locked at every third person pronoun during naturalistic story listening. We also compared the English and Chinese populations to examine whether typological differences between English and Chinese pronouns are instantiated at the brain level.

Group comparison between the activation maps for the syntax-sensitive Hobbs algorithm [49] and the discourse-sensitive ACT-R models [107] revealed distinct activation patterns, supporting a different weighting of information in English and Chinese pronoun resolution.

Given the computational components in the Hobbs and ACT-R models, we tentatively advance a functional neuroanatomy of pronoun resolution where the left IPL is involved for maintaining multiple syntactic representations, the left MTG for morphological processing, the left Precuneus for tracking multiple referents, the left AG for integrating syntactic and semantic information and the left IFG for accessing working memory.

BIOGRAPHICAL SKETCH

Jixing Li did a BA in English Language and Literature at Beijing Normal University, an MA in Linguistics at University College London and an MSc in Experimental Psychology at Oxford University. She started her PhD in Linguistics at Cornell University in 2013, and is currently doing postdoc research in the Neuroscience of Language Lab at New York University Abu Dhabi. Her research combines computational modeling with neuroimaging to examine how the human brain represents and computes syntactic and semantic information during language comprehension.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of my advisor John Hale. John's ideas inspired the work and he continues to inspire me in the future directions of my research. I am also very thankful for the freedom and belief he gave me to explore the questions I was interested in.

I thank Murielle Fabre for the stimulating discussions and for the sleepless nights we were working together before deadlines. I also thank Wen-Ming Luh for his technical support during my data collection in China.

This work also relies on the help of Yiming Yang and his team at Jiangsu Normal University who helped me conduct the fMRI experiment in China. I am grateful for the support from Jeffrey Sean Lehman Fund at Cornell University which makes my trips to China possible.

Lastly, I would like to thank my loving parents and friends, who provide unending support and inspiration in this long endeavor.

Contents

List of Tables

List of Figures

Part I

Introduction

CHAPTER 1

INTRODUCTION

1.1 Overview of the question

One fundamental aspect of human language is reference, that is, using a linguistic symbol to pick out some entity in the discourse context. This linguistic symbol is called a "referring expression", which could be a proper noun like *Mary*, a reflexive like *herself* or a pronoun like *she*. The referring expressions differ in their descriptiveness and specificity, such that proper nouns identify a unique entity in the context, while reflexives and pronouns cannot be interpreted by themselves. They are anaphors that depend their meanings on an antecedent expression. For example, in sentence (1) from the book "The Little Prince", the two characters *flower* and *the little prince* were first introduced into the discussion using proper nouns, and were later referred to using pronouns and reflexives.

(1) "My *flower* is ephemeral," *the little prince* said to *himself*, "and *she* has only four thorns to defend *herself* against the world."

Pronouns are abundant in human language and successful pronoun resolution is key to our smooth comprehension, yet the neural mechanisms underlying the process of pronoun-antecedent linking remains largely unknown. Current theories regarding pronoun resolution suggest that linguistic information such as syntactic structure and discourse salience, as well as cognitive principles of memory encoding and retrieval all constrain the interpretation of pronouns. However, previous neuroimaging studies usually tested just one or two components of theories, via manipulation of gender congruity or complexity metrics

such as distance between the antecedent and the pronoun [e.g., 44]. While these studies are helpful, they leave us without a comprehensive understanding of the computational mechanism(s) underlying pronoun resolution.

As argued by Newell [76], testing a series of binary questions about cognition might never lead to a computational understanding. He suggested that only by building comprehensive task-performing computational models can we reveal how the proposed components interact and execute the cognitive function in question. The current study is a first step towards a computational understanding of pronoun resolution in the brain. Three computational models for pronoun resolution has been evaluated against brain activity during story comprehension. Two of the models are symbolic models that instantiate the theories on pronoun resolution, and the third model is a corpus-based neural network model, which is difficult to interpret but is assumed to be more neurobiologically plausible.

We further asked the question of whether one computational model is preferred over another in different linguistic contexts. We compared the activation maps of the three models in the English and Chinese populations. English and Chinese pronouns differ in a range of typological features, hence the English-Chinese comparison is ideal for examine the cross-linguistic differences in the neural computations for pronoun resolution.

1.2 Statement of the work

This current study compared brain activity time-locked at the offset of each third person pronoun in the audiobook *The Little Prince* while English and Chinese participants listened to the story in the fMRI scanner. We correlated the

observed fMRI timeseries with processing difficulty expectations based on a syntax-sensitive Hobbs algorithm [49] and a discourse-sensitive model for pronoun resolution [107] based on the ACT-R cognitive architecture [4] (henceforth the ACT-R model for pronoun resolution). The Hobbs algorithm searches for a gender/number-matching antecedent by traversing the parsed trees in an order that respects the Binding Principles [e.g., 16]. The ACT-R model calculats the salience of the antecedent using frequency, recency and grammatical role information for all candidate antecedents. Both computational models are functionally adequate in English and Chinese with an accuracy of above 70% on the *The Little Prince* text [1].

Additionally, we tested a neural network model for coreference resolution [18] against the fMRI signals. This model encompasses a large set of features including word embeddings and some discourse features such as linear distance between the pronoun and the antecedent. Trained on the CONLL-2012 Shared Task corpus [85], this model achieved state-of-the-art results with F1 scores of 65.39 and 63.66 on the English and Chinese test data in the corpus. The high performance of the neural network model makes it a possible cognitive model for pronoun resolution.

Group comparison between the activation maps for the Hobbs and the ACT-R complexity metrics revealed distinct activation patterns for pronoun resolution in English and Chinese. Specifically, the Hobbs algorithm was associated with greater activation in a left-lateralized network including the Inferior Parietal Lobule (IPL), Precuneus, Middle Temporal Gyrus (MTG) and Middle Frontal Gyrus (MFG) in the English group, while the ACT-R model showed greater activation in the left Angular Gyrus (AG) in the Chinese group. These results

[1] Accuracy of the ACT-R model is 64% for the English text

support the hypothesis that English and Chinese speakers differ in their reliance on the syntactic and discourse factors during pronoun resolution. The neural network model was only associated with the left Superior Temporal Gyrus (STG) activation in the English group, suggesting it to be a less accurate cognitive model for pronoun resolution.

Given the components included in the Hobbs and ACT-R models, we advance a functional neuroanatomy of pronoun resolution where the left IPL is involved for maintaining multiple syntactic representations, the left MTG for morphological processing, the left Precuneus for tracking multiple referents, the left AG for integrating syntactic and semantic information and the left IFG for accessing working memory.

Compared with previous neuroimaging studies that used manipulated stimuli to investigate one or two factors in pronoun resolution such as distance [e.g., 70, 92] and gender-matching [e.g., 42] between the pronoun and the antecedent, our approach is innovative. It starts with functionally adequate computational models for pronoun resolution, and maps the components of the models onto brain structures. We also show for the first time that typological differences between English and Chinese pronouns influence the brain's algorithm for pronoun resolution in the two populations.

1.3 Structure of the dissertation

This dissertation is organized into three parts: the first part (Chapters 2 to 4) addresses the issue of pronoun resolution in English and Chinese from a linguistic, psycholinguistic and neurolinguistic point of view. The second part (Chapters 5

to 9) reviews computational approaches to pronoun resolution, describes the three computational models used in the current study and compares their model performance. The third part (Chapters 10 to 13) presents the hypotheses of the current study, the neuro-computational modeling approach, the fMRI experiment procedure, the results and discussion. A brief synopsis of each chapter is presented below.

Chapter 2 reviews the factors that have been proposed to influence pronoun resolution. We focused on two major factors: syntactic constraints and discourse preference. The syntactic constraints on pronoun resolution has been extensively discussed in the generative syntax literature [see e.g., 16, 88, 89]. These accounts suggest that pronoun interpretation is subject to a locality constraint based on syntactic structure. The discourse factors are mainly discussed in the psycholinguistic literature, where pronoun resolution is suggested to be influenced by salience of the antecedents [see e.g., 7, 39]. Both the syntactic and salience-based accounts acknowledge pronoun resolution as a complex procedure which include syntactic, semantic and discourse factors. They differ insofar as the formal syntactic theories suggest that syntactic information is accessed first before semantic and discourse-level information, whereas the psycholinguistic theories do not maintain such a hierarchy.

Chapter 3 describes the typological difference in English and Chinese. The *pro*-drop phenomenon in Chinese has been considered a typological parameter that distinguishes *topic*-prominent languages like Chinese and *subject*-prominent languages like English [66]. In addition, pronouns in spoken Chinese do not mark gender and case. Therefore, Chinese pronouns provide very few morpho-syntactic cue to help Chinese speakers identifying the antecedents. Consequently,

Chinese speakers may rely more on the salience of the antecedents during pronoun resolution compared to English speakers.

Chapter 4 surveys the neurolinguistic literature for syntactic and discourse-level processing during pronoun resolution. Brain activity associated with pronoun resolution showed a left lateralized fronto-temporal network, supporting pronoun resolution as a complex procedure that involves the integration of syntactic, semantic and discourse-level factors.

Chapter 5 briefly reviews computational approaches for pronoun resolution in the NLP literature. The models can be roughly divided into theory-driven models and corpus-driven models. Most of the early approaches to pronoun resolution were based on theoretical proposals such as syntactic constraints [e.g., 49] and discourse salience [e.g., 13, 63]. With the availability of annotated coreference corpora in the mid-1990s, corpus-based models using machine-learning methods [e.g., 77, 97] and neural network architectures [e.g., 18, 19, 64, 112] have become the current trend in the coreference resolution research.

Chapter 6 describes the Hobbs algorithm [49] and the Hobbs distance metric in detail. When applied to Chinese pronoun resolution, the gender and number agreement checker in this model was removed, which resulted in compromised model performance.

Chapter 7 reviews the modules and their corresponding brain regions proposed in the ACT-R cognitive architecture [4]. It then describes in detail the declarative modules in ACT-R, which is adapted by van Rij et al. [107] for pronoun resolution.

Chapter 8 describes the algorithm in Clark and Manning's (2016a,b) neural

coreference resolution model. The features in the neural network model include a large set of word embeddings, some discourse level features such as distance between the antecedent and the pronoun, and some word level features such as string matching and partial string matching. This model was trained on the CONLL-2012 Shared Task corpus with the training objective of successfully classifying whether a mention pair is coreferential or not.

Chapter 9 evaluates the three models' performance for third person pronoun resolution in the English and Chinese translation of the book *The Little Prince*. The Hobbs algorithm achieved higher accuracy for the English text, whereas the ACT-R model performed better for the Chinese text. The neural coreference model did not perform well on either the English or Chinese text.

Chapter 10 presents the hypotheses of the current study, namely Chinese speakers rely more on the salience of the antecedent during pronoun resolution, whereas English speakers are more sensitive to morpho-syntactic cues.

Chapter 11 describes the neuro-computational approach pioneered by Brennan et al. [12], the experimental procedure and the GLM analysis of the current study.

Chapter 12 presents brain regions associated with the presence of first, second, third person pronouns and brain regions associated with the three complexity metrics derived by the Hobbs, ACT-R and neural network models for pronoun resolution. Group comparison revealed distinct activation patterns for the Hobbs and the ACT-R metrics in the English and Chinese groups, supporting the hypotheses that English and Chinese speakers differ in their reliance on syntactic and discourse factors during pronoun resolution. The neural network model is

only associated with the left Superior Temporal Gyrus (STG) activation in the English group.

Chapter 13 discusses the functions of the brain regions observed in the current study for the Hobbs, ACT-R and neural network metrics. Based on the different functional roles of the regions suggested in the neuroimaging literature, we propose a functional neuroanatomy for pronoun resolution in English and Chinese.

Part II

Background

CHAPTER 2

THEORIES OF PRONOUN RESOLUTION

This chapter reviews linguistic and psycholinguistic theories on pronoun resolution. In formal syntactic theories, interpretation of anaphoric expressions are constrained by syntactic structures. The psycholinguistic literature, on the other hand, focuses on discourse factors that influence the interpretation of pronouns. Both the fields acknowledge that pronoun resolution is a complex procedure where syntactic, semantic and discourse factors all play a role. They differ in that the formal syntactic theories suggest that syntactic information is accessed first before semantic and discourse level information, whereas the psycholinguistic theories do not maintain such a hierarchy.

2.1 Syntactic constraints

Syntactic constraint on anaphora resolution is mainly a locality constraint. It delineates a complementary distribution of pronouns and reflexives based on syntactic configuration, although it can be overridden by cases like logophors and exceptional coreference. Later development of formal syntactic treatment on anaphora resolution suggests a modular approach that includes syntactic, semantic and discourse modules, with a hierarchy where the syntactic module is accessed first. The next section briefly reviews the development of formal syntactic theories on anaphora resolution, including Chomsky's (1981) classic Binding Theory, Reinhart and Reuland's (1993) Reflexivity theory and Reuland's (2001) Primitive of Binding framework.

2.1.1 The Binding Theory

In the generative syntax literature, the possible antecedents for a pronoun and a reflexive are constrained by syntactic structures. Under the classical Binding Theory [16], a noun phrase is divided into three types based on their ability to directly denoting entities in the real world: reflexives, pronouns and R(eferential)-expressions. Reflexives such as *herself/himself/themselves* in English or *ziji* in Chinese are anaphoric expressions that cannot directly refer to entities in the outside world and must rely on a linguistic antecedent to establish their reference. For example, *himself* and *taziji* in Sentence (1) and (2) can only refer back to *Bill* and *Lisi*, respectively. Third person pronouns like *she/he/it/they* in English or *ta/tamen* in Chinese can be anaphoric or deictic: *him* and *ta* in (3) and (4) can either refer back to *John/Zhangsan* or some other person in the discourse context. R-expressions are other noun phrases that do not need antecedents, such as *John/Bill* and *Zhangsan/Lisi*.

(1) John$_i$ thinks that [$_{IP}$Bill$_j$ always criticizes himself$_{*i/j/*k}$].

(2) Zhangsan$_i$ juede [$_{IP}$Lisi$_j$ zongshi piping taziji$_{*i/j/*k}$].

Zhangsan$_i$ think Lisi$_j$ always criticize himself$_{*i/j/*k}$.

(3) John$_i$ thinks that [$_{IP}$Bill$_j$ always criticizes him$_{i/*j/k}$].

(4) Zhangsan$_i$ juede [$_{IP}$Lisi$_j$ zongshi piping ta$_{i/*j/k}$].

Zhangsan$_i$ think Lisi$_j$ always criticize him$_{i/*j/k}$.

(5) John$_i$ thinks that [$_{IP}$he$_{i/j}$ always criticizes Bill$_{*j/k}$].

(6) Zhangsan$_i$ juede [$_{IP}$ta$_{i/j}$ zongshi piping Lisi$_{*j/k}$].

Zhangsan$_i$ think he$_{i/j}$ always criticize Lisi$_{*j/k}$.

Three structural principles govern the possibility and location of an antecedent for a noun phrase:

Principle A: An anaphor is bound in its local domain.

Principle B: A pronominal is free in its local domain.

Principle C: An R-expression is free.

A "local domain" can be roughly defined as the smallest IP or NP containing the predicate that assigns the theta roles, the complements to which the internal theta roles are assigned, and the subject to which the external theta role is assigned. [1] In Sentence (1)-(4), the "local domain" is IP, therefore, Principle A requires the reflexive *himself/taziji* in (1)-(2) to be coindexed with the subject of the IP *Bill/Lisi*, and Principle B requires the pronoun *him/ta* in (3)-(4) to not refer to *Bill/Lisi*. Principle C also successfully rules out *he/ta* and *Bill/Lisi* to corefer in (5)-(6), as the R-expression *Bill/Lisi* cannot be bound.

2.1.2 Reflexivity

One crucial implication of Principle A and B is that reflexives and pronouns are in a complementary distribution, namely, reflexives do not occur in positions where pronouns are used. This seems to hold in many canonical examples (see (1)-(6)). However, there are some notable exceptions as exemplified in (7)–(12), where both the pronoun *him/ta* and the reflexive *himself/taziji* are grammatically acceptable.

(7) John$_i$ saw a picture of himself$_i$/him$_i$.

[1] Although the exact definition of "local domain" is much more complicated.

(8) Zhangsan$_i$ kanjian le taziji$_i$/ta$_i$ de huaxiang.

Zhangsan$_i$ saw himself$_i$/him$_i$ DE picture.

(9) John$_i$ said [$_{IP}$that the queen$_j$ invited Bill$_k$ and himself$_i$/him$_i$ to tea].

(10) Zhangsan$_i$ shuo [$_{IP}$nvwang$_j$ yaoqing le Lisi$_k$ he taziji$_i$/ta$_i$].

Zhangsan$_i$ said queen$_j$ invited Lisi$_k$ and himself$_i$/him$_i$.

(11) I know what Mary$_i$ and John$_j$ have in common. Mary$_i$ adores him$_j$ and John$_j$ adores him$_j$ too.

(12) Wo zhidao Zhangsan$_i$ he Lisi$_j$ you shenme gongtongdian. Zhangsan$_i$ xihuan ta$_j$, Lisi$_j$ ye xihuan ta$_j$.

I know Zhangsan$_i$ and Lisi$_j$ have what in common. Zhangsan$_i$ adores him$_j$, Lisi$_j$ too adores him$_j$.

The numerous violations to the complementary distribution of pronouns and reflexives have led researchers to re-examine Chomsky's Binding Theory in its classical form. Reinhart and Reuland [88], for example, noted that the critical element for the Binding Theory is whether the pronoun/reflexive and its antecedent are arguments of the same predicate. Reflexives are able to reflexivize the predicate, that is, they impose identity on two arguments of a predicate. Reinhart and Reuland [88] reformulated the Binding Principles as Condition A and B:

Condition A: A reflexive-marked predicate is reflexive.

Condition B: A reflexive predicate is reflexive marked.

Condition A ensures that reflexives and their co-arguments refer to the same entity as in "Bill$_i$ always criticizes himself$_i$", and Condition B rules out the use of

pronouns in "Bill$_i$ always criticizes him$_{*i}$" since it states that if the co-arguments of a predicate is co-indexed, it should be marked by a reflexive. For cases in (7)-(10), both reflexives and pronouns are acceptable because their antecedents are not co-arguments of a same predicate. In (7)-(8) *John*$_i$ is an argument of the verb *saw* and *himself*$_i$/*him*$_i$ is an argument of the preposition *of*, in (9)-(10) *John*$_i$ is an argument of *said* and *himself*$_i$/*him*$_i$ is an argument of *invite*, thus they are not excluded by either Condition A or B. The reflexives used in cases (7)-(10) are referred to as logophors.

Reinhart and Reuland therefore proposed a modular approach on the interpretations of reflexives: for reflexives whose antecedents are co-arguments of the same predicate as in (1)-(2), the anaphoric dependency is encoded in syntax; for logophoric reflexives in (7)-(10), however, the interpretation of the reference involves an inference based on meaning and appropriateness of discourse context.

Another motivation for a modular approach for the interpretation of pronouns comes from the "exceptional coreference" cases in (10)-(11), first observed in Evans [24]. According to Grodzinsky and Reinhart [38], (10)-(11) have two interpretations at the semantic level: (a) *him* is interpreted as a variable and must be linked to the suitable binder *John*; (b) *him* is interpreted referentially and is insensitive to the specific value of a, thus the interpretation where a = *John* is included as well.

a. John λx (x adores x).

b. John λx (x adores a).

In normal cases, the exceptional coreference reading where a = *John* as in

"John$_i$ adores him$_{*i}$" is blocked because there is a direct variable binding option which forces the use of the reflexive *himself*. *a = John* is only allowed where the coreference interpretation yields a different interpretation from variable binding. In cases (10)-(11), the coreference reading "Mary$_i$ adores him$_j$ and John$_j$ adores him$_j$" states a property that is shared by Mary and John, different from the bound variable reading where John adores himself. This interpretive condition on the coreference option is formulated as "Rule I" in Grodzinsky and Reinhart [38]:

Rule I: Intrasentential Coreference:

NP A cannot corefer with NP B if replacing A with C, C a variable A-bound by B, yields an indistinguishable interpretation.

Rule I therefore reflects a division of labor within the linguistic system: encoding dependency relation via variable binding and establishing coreference relation via discourse context. Reinhart [87] suggested that a bound variable reading is preferred over a coreference reading for independent reasons like "early closure" of an open expression. That is, variable binding through syntactic encoding is more readily accessible, or "less costly" than establishing coreference by accessing discourse storage. This modular view on the interpretation of pronouns and reflexives has been further explored in Reuland's Primitives of Binding framework.

2.1.3 The Primitive of Binding framework

The architecture of the Primitives of Binding (POB) framework consists of a syntactic, semantic and discourse module. The syntactic module is specialized for establishing dependencies between the reflexives and their antecedents via

A-Chain formation. According to Reuland [89], an A-Chain can only be formed if the anaphoric element is deficient in contents, namely, it lacks a fully specified set of ϕ-features like person, gender, number, and Case. This led to the locality constraint stated in Binding Principle A on reflexive binding. Pronouns and logophoric reflexives, on the other hand, contain fully specified ϕ-features, hence they cannot be bound with their antecedents in the syntactic module via A-Chain. Instead, they are either bound via variable binding in the semantic module, or co-indexed via coreference in the discourse module (see Figure 2.1 for a schematic illustration of the POB framework).

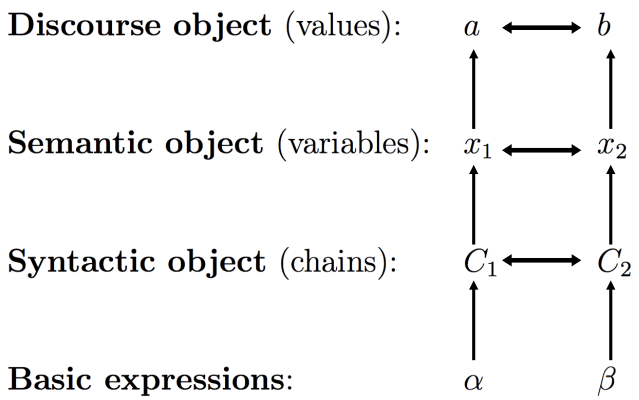


Figure 2.1: schematic illustration of the syntactic, semantic and discourse module in the POB framework. Reflexives are bound via syntactic Chains in the syntactic module; pronouns are bound either by variable binding in the semantic module (or the Conceptual-Intentional interface), or by coreference in the discourse module. The syntactic operation precedes the semantic operation, which in turn precedes the discourse operation.

A central claim of the POB framework is that the syntactic, semantic and discourse modules are accessed sequentially in the interpretative steps of anaphoric expressions, resulting in increased processing load. Therefore, according to the "economy principle", A-Chain formation in the syntactic module would block

variable binding in the semantic module, which would in turn block coreference in the discourse module. This is similar to Grodzinsky and Reinhart's (1993) Rule I where variable binding blocks coreference reading. Division of labor among these sub-systems of grammar leads to complementarity of reflexive and pronoun binding.

Thus far, I have briefly discussed the constraints on anaphora resolution from a syntactic point of view. As suggested in the Reflexivity theory and the POB model, the relatively complex pattern of the use of anaphora calls for a modular approach where syntactic, semantic and discourse subsystems interact in real time anaphora processing. It is not concerned in these formal linguistic theories how coreference relation is established in the discourse module, yet this question is central to pronoun resolution in the cognitive and psycholinguistic literature. In the next section I briefly reviews the most prominent theories on anaphora resolution from a discourse point of view.

2.2 Discourse preference

From a discourse point of view, anaphoric expressions refer to a highly salient entity in the context. Different factors interact in parallel and all contribute to the salience of entities in the discourse context. This section reviews two most prominent theories of the salient-based account on pronoun resolution.

2.2.1 The Centering Theory

The Centering Theory [39] concerns the perceived coherence of utterances within a discourse segment and the choice of referring expression. It claims that certain entity mentioned in an utterance is more "central" than others, which leads the speaker to use different types of referring expressions. In the Centering framework, entities serving to link an utterance to other utterances in the discourse segment is referred to as "centers". Each utterance (U) has a set of forward-looking centers (C_fs) and a single backward-looking center (C_b). The C_b of utterance U_{n+1} connects with one of the C_fs in the previous utterance U_n. For example, in the following discourse segment, the set of C_fs in (13c) contains *John* and *Bill*, and the C_b of (13c) is *John*, which connects to one of the C_fs ({ *John*, *Bill* }) in (13b).

(13) a. John has been having a lot of trouble arranging his vacation.

b. He cannot find anyone to take over his responsibilities.

(C_b = John; C_f = John, he = John)

c. He called up Bill yesterday to work out a plan.

(C_b = John; C_f = John, Bill, he = John) **CONTINUATION**

d. Bill has annoyed him a lot recently.

(C_b = John; C_f = John, Bill) **RETAINING**

e. He called John at 5 AM on Friday last week. (he = Bill)

(C_b = Bill; C_f = John, Bill) **SHIFTING**

The elements of $C_f(U_n)$ are ranked to reflect their relative prominence in U_n. The more highly ranked an C_f is in U_n, the more likely it is the C_b in U_{n+1}. A

number of factors affect the ranking of C_fs, including grammatical roles, word order (especially fronting), clausal subordination, and lexical semantics, etc. The most influential factor discussed in Grosz et al. [39] is grammatical role, in particular: SUBJECT > OBJECT > OTHERS. Thus the subject of the U_n is more likely to be the C_b of U_{n+1}. In example (13), the pronoun *he* in (13b) is in a subject position, thus it is the C_b of (13c). Similarly, *Bill* in (13d) is in a subject position, and it is therefore the C_b of (13e).

Another important notion in the Centering framework is the transition relations between pairs of utterances, which influence the coherence of the discourse. There are three transition relations: Continuation, Retaining and Shifting. In the Continuation transition, C_b of current utterance is the C_b of the previous sentence, and it is also the mostly highly ranked element in the C_fs of the current sentence, so it is likely to be the C_b of the next sentence. In the Retaining transition, C_b of current utterance is the C_b of the previous sentence, but it is not the mostly highly ranked element in the C_fs and is unlikely to be the C_b of the next sentence. In the Shifting transition, C_b of current utterance is not the C_b of the previous sentence (See (13c)-(13e)).

CONTINUATION: $C_b(U_{n+1}) = C_b(U_n)$, and $C_b(U_{n+1}) = C_b(U_{n+2})$.

RETAINING: $C_b(U_{n+1} = Cb(Un)$, but $C_b(U_{n+1}) \neq C_b(U_{n+2})$

SHIFTING: $C_b(U_{n+1} \neq Cb(Un)$.

For a discourse segment to be coherent, sequences of Continuation are preferred over sequences of Retaining, which are in turn preferred over sequences of Shifting. Frequent Shifting, as exemplified in (14) and (15), leads to a lack of discourse coherence and substantially affects the processing demands made

upon a hearer during discourse comprehension.

14 a. Susan is a fine friend.

b. She gives people the most wonderful presents.

(C_b = Susan; C_f = Susan)

c. She just gave Betsy a wonderful bottle of wine.

(C_b = Susan; C_f = Susan,Besty) **CONTINUATION**

d. She told her it was quite rare.

(C_b = Susan; C_f = Susan,Besty) **CONTINUATION**

e. She knows a lot about wine.

(C_b = Susan; C_f = Susan) **CONTINUATION**

15 a. Susan is a fine friend.

b. She gives people the most wonderful presents.

(C_b = Susan; C_f = Susan)

c. Besty was given a wonderful bottle of wine.

(C_b = Susan; C_f = Besty) **RETAINING**

d. Susan told her it was quite rare.

(C_b = Besty; C_f = Susan,Besty) **SHIFTING**

e. Besty knows a lot about wine.

(C_b = Susan; C_f = Besty) **SHIFTING**

2.2.2 Accessibility of reference

Arnold [7] proposed a similar salience-based account for pronominalization. She suggested that expressions for referents in the discourse contexts fall along a hierarchy of explicitness [see 1, 6, 40, for a review], ranging from semantically rich expressions ("the house with a red door") to shorter terms ("the house") to pronouns ("it") and even zeros (dropped pronouns in *pro*-drop languages like Chinese and Spanish). Preferences for these forms of expressions depend on the "accessibility" of the referents in the discourse context: more accessible referents tend to be less explicit; by contrast, more explicit expressions are used for less salient references.

The reason for using less-explicit forms for more accessible referents could be that less-specific forms are more efficient for communication, as embodied in Grice's maxim of quantity: Make your contribution as informative as required, but not more so [36]. A second explanation is that the referential form is a "marker" for the discourse status of the referent, which helps the listeners to identify the referent in their mental representation. This idea is supported by the "repeated name penalty" phenomenon where reading time was slowed when a repeated name was used for a highly accessible referent [e.g., 35].

The accessibility of a referent, according to Arnold [7], is mainly influenced by four discourse properties: givenness, recency, syntactic prominence and semantic prominence. Givenness concerns whether the referent has been mentioned before in the discourse context. Pronouns are generally reserved for already-mentioned referents, whereas new things are introduced with more explicit expressions. Recency refers to how recent the referent occurs in the discourse; more recently-mentioned information is more accessible and more likely to be pronominalized.

Syntactically prominent items, such as subject of a clause, or focus of cleft, are perceived as more accessible. In addition, the semantic role of an entity also influences the entity's accessibility. For example, in transitive events with Stimulus and Experiencer roles, pronouns are preferred to refer to the Stimulus role (see (16)).

(16) a. Experiencer-Stimulus: Hannah$_i$ admired Laura$_j$ enormously because she$_j$...

b. Stimulus-Experiencer: Hannah$_i$ impressed Laura$_j$ enormously because she$_i$...

Arnold [7] further suggested that accessibility could be modeled as gradient activation of discourse representations. The four discourse properties affects the activation of a representation, for example, subjecthood will increase the activation of a referent. If the activation exceeds a particular threshold, pronouns are generally preferred.

Both the Centering Theory and the Accessibility theory suggested some sort of mechanism that selects the most salient entity in the discourse context as the referent of a pronoun. This mechanism likely involves basic cognitive operations such as memory retrieval, where a prominent antecedent is actively maintained in focal attention. Factors influence the selecting mechanism proposed by the two theories include grammatical role, transition relations, givenness, recency, syntactic and thematic prominence. The next section briefly reviews some experimental studies that support the involvement of the proposed factors during pronoun resolution.

2.2.3 Psycholinguistic evidences

The Centering Theory has been tested by Gordon et al. [35] in a number of self-paced reading experiments. They introduced a prominent entity (C_b) and a less prominent entity in a short passage and found that reading time significantly increased when the prominent entity is not pronominalized but repeated. They also showed this repeated-name penalty for C_b only in the grammatical subject position, confirmed the basic notion in the Centering Theory that there is only one C_b in an utterance, and that grammatical subject ranks the highest among the C_fs.

First-mentioned referent is also more likely to be the antecedent of a pronoun. Järvikivi et al. [57], using a visual-world eye-tracking paradigm, presented sentences containing an ambiguous pronoun that referred to either the subject or the object of a previously presented sentence in an SVO or OVS order. Participants' eye movements were monitored while they looked at pictures representing the two possible antecedents of each pronoun. The results revealed that they used both order-of-mention and grammatical role information to resolve ambiguous pronouns.

Recency of the antecedent on pronoun resolution is examined in a written corpus in Arnold [7]. The data showed that over 90% of the pronouns have antecedents within the same clause or one clause before. By contrast, only about 30% of all types of references refer to an entity in the current or the previous clause.

The role of syntactic clefting in pronoun resolution is examined by Foraker and McElree [30] using self-paced reading and eye-tracking tasks. They pre-

sented sentences with clefted antecedents like “It was the new foreman who unrolled the latest blueprint. He squinted at the lines of the paper.”, and found that although clefting did not affect the speed of accessing the antecedent, it increased the likelihood of retrieving the antecedent, suggesting that clefting made antecedent representations more distinctive in working memory, hence more available for subsequent discourse operations.

CHAPTER 3

PRONOUN RESOLUTION IN ENGLISH AND CHINESE

One major difference of English and Chinese pronouns is that Chinese allows pronouns to be deleted in finite sentences at both the subject and object positions. This phenomenon is called *pro*-drop in generative syntax and has been considered a typological parameter that distinguishes "topic-prominent" languages like Chinese and "subject-prominent" languages like English [66]. In addition, pronouns in spoken Chinese do not mark gender and Case. Therefore, Chinese pronouns in general provide little morpho-syntactic cues to help Chinese speakers identify the correct antecedent. Consequently, compared to English speakers, Chinese speakers may rely more on discourse information during pronoun resolution. In the rest of the section I briefly reviewed the syntactic approach to the *pro*-drop phenomenon in Chinese and its typological implication. I then discussed the possible consequences of the typological difference on the English and Chinese comprehenders' cognitive states during pronoun resolution.

3.1 Typological differences

In English, pronouns cannot be omitted in the subject or object position of a tensed clause, even though the reference of the omitted pronoun is clear from the context. On the contrary, Chinese can have a null pronoun as the subject or object of a tense clause in appropriate contexts (see (9) and (10), data from Huang [51]). Besides Chinese, Japanese and Korean also allow null subjects and objects, whereas some European languages like Italian and Spanish allow null subjects but not null objects. These languages are therefore called *pro*-drop or null subject

languages.

(9) Speaker A: Did John see Bill yesterday?

Speaker B:
a. Yes, he saw him.
b. *Yes, *e* saw him.
c. *Yes, he saw *e*.
d. *Yes, *e* saw *e*.
e. *Yes, I guess *e* saw *e*.
f. *Yes, John said *e* saw *e*.

(10) Speaker A: Zhangsan kanjian Lisi le ma?
Zhangsan see Lisi LE Q?
"Did Zhangsan see Lisi?"

Speaker B:
a. Ta kanjian ta le.
He see he LE.
"He saw him."
b. *e* kanjian ta le.
"[He] saw him."
c. Ta kanjian *e* le.
"He saw [him]."
d. *e* kanjian *e* le.
"[He] saw [him]."
e. Wo cai *e* kanjian *e* le.
I guess see LE.
"I guess [he] saw [him]."
f. Zhangsan shuo *e* kanjian *e* le.
Zhangsan say see LE.
"Zhangsan said that [he] saw [him]."

Taraldsen [101] and Rizzi [90], among others, argued that null subjects are allowed in Italian and Spanish because the verb-subject agreement marking on

the verb is rich enough to recover the content of the missing subject; null objects are not allowed because there is no verb-object agreement in these languages. For English and French, the verb-subject agreement system is somewhat degenerate, and the agreement marking on the verb is too meager to identify the omitted subject. Thus *pro*-drop is not allowed in English and French. This account, however, cannot explain the even more "radical" *pro*-drop in Chinese, Japanese and Korean, which have neither verb-subject or verb-object agreement marking on the verb.

Li and Thompson [66] suggested that the distribution of null pronouns in different languages could reflect a more general typological parameter: the "topic-prominent" vs. the "subject-prominent" parameter. Chinese is a topic-prominent language with a "topic-comment" sentence structure (see (11)), whereas English is a subject-prominent language that has a "subject-predicate" structure. Subject-prominent languages must have subjects, as described by the Extended Projection Principle in Chomsky [17]; topic-prominent languages like Chinese do not require structural subjects, hence do not have pleonastic elements like *it* and *there* in English. It then follows naturally that the topic-comment structure allows independent sentences to drop the topic that can be identified in the context. Tsao [102] suggested that Chinese is "discourse-oriented" and has a "Topic NP deletion" rule which allows the topic to be deleted if it is the same with the topic in the preceding sentence (see (12)); English, however, is sentence-oriented and lacks this topic-chain interpretation rule.

(11) Neichanghuo, xingkui xiaofangdui lai de zao.

that fire, fortunately fire-brigade come COMP early.

"That fire, fortunately the fire-brigade came early."

(12) Zhongguo difang hen da, *e* renkou hen duo, *e* tudi hen feiwo, *e* qihou ye hen hao, *e* women dou hen xihuan.

China place very big, *e* population very many, *e* land very fertile, *e* climate too very good, *e* we all very like.

"As for China, (its) land area is very large; (its) population is very big; (its) land is very fertile; (its) climate is also very good. We all like (it)."

Huang and Barry [52] proposed that there is an uninterpretable topic feature [*u*Top] that is pending for valuation at C in Chinese sentences. Null topic is licensed in Chinese via the checking of the [*u*Top] feature at C. When there is an overt topic with the interpretable [*i*Top] feature, it directly merges to CP and checks the [*u*Top] feature at C; when the topic is omitted, the uninterpretable [*u*Top] feature probes into its domain to find an appropriate *pro* and attracts it. This movement is blocked by any island on the path. For example, the *pro* cannot move out of the complex DP in (13b) to check the [*u*Top] feature at the matrix C, thus (13b) is ungrammatical; (13a) is grammatical because the overt Topic *Zhangsan* directly merges with the matrix C.

Merge: [$_{CP}$ TopicP$_{[iTop]}$ C$_{[uTop]}$ ... [$_{IP}$ ... *pro* ...]].

Move: [$_{CP}$ *pro* C$_{[uTop]}$ ... [$_{IP}$ ... *e* ...]].

*[$_{CP}$ *pro* C$_{[uTop]}$ ... [$_{island}$ ... *e* ...]].

(13) a. Xianzai wo lai shuoshuo Zhangsan$_i$. [$_{CP}$ Zhangsan$_i$, [$_{DP}$ xuduo [$_{CP}$ *pro*$_i$ xie de] shu] dou hen changxiao].

me come talk Zhangsan$_i$. Zhangsan$_i$, many *pro*$_i$ write DE book all very sell-well.

"Now let me talk about Zhangsan$_i$. Zhangsan$_i$, many books that [he$_i$] wrote sell well."

b. *Xianzai wo lai shuoshuo Zhangsan$_i$. [$_{CP}$ [$_{DP}$ xuduo [$_{CP}$ *pro$_i$* xie de] shu] dou hen changxiao].

*Now me come talk Zhangsan$_i$. Many *pro$_i$* write DE book all very sell-well.

*"Now let me talk about Zhangsan$_i$. Many books that [he$_i$] wrote sell well."

3.2 Implications for pronoun resolution in English and Chinese

The syntax approach for *pro*-drop and the "topic-prominent" vs. the "subject-prominent" parameters offer an important insight on the neural mechanisms of pronoun resolution in English and Chinese. Since "topic" is more of a discourse notion and "subject" is more related to syntactic analysis, it is hypothesized that English and Chinese speakers use different mechanisms to resolve the reference of pronouns: English speakers rely more on structural and morpho-syntactic analysis, whereas Chinese speakers are more sensitive to discourse information.

In addition, for English speakers, gender, number and case markings on the pronouns all provide cues to search for the correct antecedent. However, Chinese pronouns do not mark gender in their spoken forms, and although there is plural marking "-men" on pronouns, there is no plural marking on NPs in Chinese. Therefore, both gender and number information is absent

during pronoun resolution in Chinese. Moreover, Chinese pronouns can even be omitted. This leads to the question of how Chinese speakers recover the correct antecedent of the pronoun. One hypothesis is that Chinese speakers always use constructions where the referent of the pronoun is the most salient entity in the discourse context. For example, in (14) where *John* and *Mary* are equally salient in the discourse context, English speakers can use gender information to figure out the correct referent of the pronoun but Chinese speakers cannot, so that sentence like (15b) following (15a) in Chinese is pragmatically illegitimate. In this situation, Chinese speakers either use full names to avoid ambiguity as in (15c), or change the constructions in (15a) to (16a) to make the referent more salient (e.g., in a subject position) in the discourse context.

(14) a. John$_i$ and Mary$_j$ are good friends.

b. He$_i$ gave her$_j$ a pet hamster on her birthday.

(15) a. Xiaoming$_i$ he Xiaohong$_j$ shi hao pengyou.

Xiaoming$_i$ and Xiaohong$_j$ are good friends.

*b. ta$_i$ zai ta$_j$ shengri de shihou song le ta$_j$ yi zhi cangshu.

He$_i$ on her$_j$ birthday gave her$_j$ a pet hamster.

c. Xiaoming$_i$ zai ta$_j$ shengri de shihou song le ta$_j$ yi zhi cangshu.

Xiaoming$_i$ on her$_j$ birthday gave her$_j$ a pet hamster.

(16) a. Xiaoming$_i$ shi Xiaohong$_j$ de hao pengyou.

Xiaoming$_i$ is Xiaohong$_j$'s good friend.

b. ta$_i$ zai ta$_j$ shengri de shihou song le ta$_j$ yi zhi cangshu.

He$_i$ on her$_j$ birthday gave her$_j$ a pet hamster.

The current study examines the hypothesis that English and Chinese speakers differ in their neural mechanisms for pronoun resolution. Specifically, we compared whether a syntax-sensitive model or a discourse-sensitive model on pronoun resolution fits the brain activity while English and Chinese participants listen to a same story that contains hundreds of third person pronouns. If the syntax-sensitive model fits better with the English speakers' brain activity and the discourse-sensitive model fits better with the Chinese speakers' brain activity, then the hypothesis that typological difference of pronouns influences the neural mechanisms of pronoun resolution in English and Chinese is supported. The syntax-sensitive and discourse-sensitive models are also compared to the corpus-based neural coreference model to examine whether pronoun resolution could be viewed as an emergent property from the brain's responses to the task it is presented. Chapter 10 describes the hypotheses and predictions of the current study in detail.

CHAPTER 4

NEUROLINGUISTIC EVIDENCES

4.1 Mechanisms for pronoun resolution

Most previous neurocognitive studies on pronoun resolution used the event-related brain potentials (ERPs) technique, which provides information about neural activity with fine temporal resolution. Yet to date there are only a few neuroimaging studies on pronouns, with no consensus on what brain regions are responsible for pronoun resolution. The remainder of this section briefly reviews the findings from previous ERP and fMRI studies on the neural mechanisms of pronoun resolution. The evidences suggest that both syntactic and discourse-level processing are involved in pronoun comprehension.

4.1.1 Evidence for syntactic processing

Violation of syntactic constraints on pronoun resolution was found to induce a larger P600, which has been traditionally associated with syntactic analysis or reanalysis. For example, a gender or number mismatch between an anaphor and the antecedent in the sentence, as in "The hungry guests helped themselves/*himself to the food" and "The successful woman congratulated herself/*himself on the promotion", resulted in an increased P600 effect relative to the controls [80]. This effect was also observed when the antecedent has a stereotypical gender, as in "The doctor prepared himself/herself for the operation" [81]. In addition, in languages where nouns have grammatical gender (e.g.,

in French, "table (table)" is feminine and "chaise (chair)" is masculine), violation of grammatical gender agreement also elicited a larger P600 [94].

But the P600 effect does not simply index a gender/number mismatch between an anaphor and the subject. Harris et al. [45] compared sentences which contain either an anaphor or a logophor. An anaphor, like "himself" in "The pilot's mechanics brow-beat *himself after the race", has to be coindexed with the matrix subject "the pilot's mechanics". Thus "himself" is not allowed in this sentence as it disagrees with the antecedent in number. A logophor, on the other hand, is not in the argument position of a verb and can refer to an embedded subject. For example, in the sentence "The pilot's mechanics brow-beat Paxton and himself after the race", the logophor "himself", although also disagrees with the matrix subject in number, is grammatical as it refers to the embedded subject "pilot". The ERP results showed that the anaphor "himself" induced a large P600 effect, while the logophor "himself" did not. This suggested that participants were sensitive to the syntactic structure of a sentence, not just the superficial gender/number mismatch.

For sentences where a pronoun disagrees in gender with an antecedent but could refer to a third, unmentioned person outside the context, such as "The aunt heard that he had won the lottery", the P600 effect was also observed, although the sentence could be grammatically correct [78, 80]. Osterhout and Mobley [80] further grouped the participants according to their judgments on the acceptability of these sentences, and they found the P600 effect for participants who judged the sentences to be unacceptable, but no such effect was found for participants who accepted these sentences. This suggested that syntactic analysis does play a role in the processing of "referentially failing pronouns".

4.1.2 Evidence for discourse processing

Evidence supporting discourse-level processing in pronoun resolution comes from studies that do not involve ill-formed sentences. Salience-based account on pronoun resolution suggest that the most salient entity is the correct antecedent of the pronoun (see Section 2.2). Under these theories, when encountering a pronoun, the comprehender matches it to the most salient entity in the current working memory buffer; when the antecedent is less salient, a full expression such as a repeated name is used [30]. This leads to the prediction that difficulty of pronoun resolution decreased by antecedent prominence, while difficulty for repeated name resolution increased by antecedent prominence. Swaab et al. [100] tested this prediction in an ERP experiment where participants read sentences containing a prominent antecedent and a pronoun/repeated name, such as "John went to the store so that John/he could buy some candy", and sentences containing a less prominent antecedent and a pronoun/repeated name, such as 'John and Mary went to the store so that John/he could buy some candy". They found that repeated names referring to a prominent antecedent elicited a larger N400, which is generally associated with difficulty of semantic integration. This supports Gordon and Hendrick's [34] theory that the primary purpose of names is to introduce entities into a discourse model, thus using repeated names for coreference requires subsequent integration of the new entity with an already existing entity in the discourse model. However, for pronouns co-indexed with a less prominent antecedent, no similar N400 effect was found, suggesting that pronoun resolution may not be solely determined by antecedent prominence.

Studies on referential ambiguity also support the salience-based account of pronoun resolution. Since two antecedents could be equally salient in the

discourse, referential ambiguity may arise and cause a processing load for the comprehender. van Berkum et al. [104] first examined the processing consequences of referential ambiguity in an ERP study. They presented participants with two stories containing either one or two possible antecedents, and they found that the referentially ambiguous word in the two-referent condition elicited a sustained, negative ERP, which they called the "Nerf" effect. The Nerf effect was replicated in van Berkum et al. [105] where the stimuli was presented in the auditory modality. Since the Nerf effect is different from an either an N400 or a P600 effect, van Berkum et al. [104] suggested that the processing cost induced by referential ambiguity is different from a semantically or syntactically problematic word. Nieuwland and Van Berkum [78] further investigated individual differences in the processing of ambiguous pronouns, and they found a larger Nerf effect for participants with a higher reading span, suggesting that ambiguous pronoun resolution is related to working memory capacity.

4.1.3 Evidence for syntax-discourse interaction

To disentangle the effect of syntactic violation and discourse salience on pronoun resolution, Hammer et al. [43] manipulated the syntactic gender matching and the distance between the antecedent and the pronoun using German sentences:

Short distance	congruent:	Der Apfel$_{mas}$ ist süß, weil er$_{mas}$ reif ist.
	incongruent:	Der Apfel$_{mas}$ ist süß, weil sie$_{fem}$ reif ist.
		"The apple is sweet, because he/she (it) ripe is."
Long distance	congruent:	Der Apfel$_{mas}$ ist sehr saftig und süß, weil er$_{mas}$ reif ist.
	incongruent:	Der Apfel$_{mas}$ ist sehr saftig und süß, weil sie$_{fem}$ reif ist.
		"The apple is very juicy and is sweet, because he/she (it) ripe is."

They argued that if syntactic processing and discourse salience are independent of each other, they should expect a P600 effect for the incongruent sentences compared to the congruent sentences, and an increased LAN component for long distance compared to short distance conditions. Alternatively, if discourse salience interacts with syntactic processing, then there should be a difference in the effect size of the P600 in long distance compared to short distance conditions. The ERP results supported the interactive view: they found a P600 effect for short incongruent sentences but not for the long incongruent sentences, suggesting that syntactic gender violation cannot be detected anymore if the antecedent and the pronoun are distant.

4.2 Brain regions involved in pronoun resolution

The previous section reviews EEG studies supporting both syntactic and discourse-level processing during pronoun resolution. This section summarizes brain regions that have been reported in the fMRI literature that correlate with pronoun resolution. Ideally, a meta-analysis should be conducted to list the regions common to pronoun resolution under different experimental tasks, yet the number of fMRI studies on pronoun resolution is too small for such an analysis.

van Berkum et al. [106] compared the BOLD responses when participants read sentences containing a "referentially failing pronoun" (e.g., "Rose told Emily that *he* had a positive attitude towards life.") or a coherent pronoun (e.g., "Ronald told Emily that *he* had a positive attitude towards life."). The results showed that referentially failing pronouns were associated with increased activation in the

medial parietal regions and bilateral inferior parietal regions, possibly reflecting morpho-syntactic processing.

Hammer et al. [42] explored grammatical gender mismatch in pronoun resolution by manipulating the syntactic gender matching between the antecedent and pronouns using German sentences. The results showed that incongruency of syntactic gender between the pronoun and its antecedent activated the bilateral Inferior Frontal Gyrus (IFG), the left Medial Frontal Gyrus (MFG) and the bilateral Supramarginal/Angular Gyrus compared to congruent pronoun-antecedent pairs. Hammer et al. [44] further investigated the possible interactions between gender and distance (the example sentences are shown below), and they reported a fronto-temporal network including the bilateral IFG, the Superior Temporal Gyrus (STG) and posterior Middle Temporal Gyrus (pMTG) for long distance conditions, with the pMTG additionally driven by syntactic gender violation. They suggested that the temporal regions are sensitive to the morpho-syntactic information of the antecedents, since long distance between the antecedent and the pronoun increased the overall syntactic complexity of the sentence.

person	short distance	congruent:	Der Häuptling$_{MALE/mas}$ ist kriegerisch, weil **er**$_{MALE/mas}$ gewinnen will.
		incongruent:	Der Häuptling$_{MALE/mas}$ ist kriegerisch, weil **sie**$_{FEMALE/mas}$ gewinnen will.
			"The chief is matial, because he/she win want."
person	long distance	congruent:	Der Häuptling$_{MALE/mas}$ greift bald an und ist kriegerisch, weil **er**$_{MALE/mas}$ gewinnen will.
		incongruent:	Der Häuptling$_{MALE/mas}$ greift bald an und ist kriegerisch, weil **sie**$_{FEMALE/mas}$ gewinnen will.
			"The chief attacks soon and is matial, because he/she win want."
thing	short distance	congruent:	Der Apfel$_{mas}$ ist süβ, weil **er**$_{MALE/mas}$ reif ist.
		incongruent:	Der Apfel$_{mas}$ ist süβ, weil *sie*$_{FEMALE/fem}$ reif ist.
			"The apple is sweet, because he/she ripe is."
thing	long distance	congruent:	Der Apfel$_{mas}$ ist sehr saftig und ist süβ, weil **er**$_{MALE/mas}$ reif ist.
		incongruent:	Der Apfel$_{mas}$ ist sehr saftig und süβ, weil *sie*$_{FEMALE/fem}$ reif ist.
			"The apple is very juicy and is sweet, because he/she ripe is."

Matchin et al. [70] also examined the effect of distance but with the backward anaphora/filler-gap dependencies contrast (the example sentences are shown below). They observed the bilateral Anterior Temporal Lobes (ATLs), the bilat-

eral Angular Gyrus, and the left Precuneus activation during the processing of backward anaphora compared to wh-fillers. Consistent with Hammer et al.'s [44] finding, they also found significant left IFG activation for long distance between antecedent and backward anaphora. The right Superior Temporal Sulcus (STS) and the Supplementary Motor Areas (SMAs) were also associated with long distance compared to short distance for backward anaphora.

wh-filler	short distance:	Which song_1 did the band play_1 at the concert [that ended early]?
	long distance:	Which song_1 did the band [that won the contest] play_1 at the concert?
backward anaphora	short distance:	Because he_1 extinguished the flames, the fireman_1 saved the resident [that arrived later].
	long distance:	Because he_1 extinguished the flames [that burned all night long], the fireman_1 saved the resident.

Santi and Grodzinsky [93] compared the activation maps for a null pronoun *PRO*, a parasitic-gap and a wh-trace in sentences such has "[Which paper] did the tired student submit [wh-trace] after reviewing [parasitic gap/PRO]?". The results showed the right Middle Frontal Gyrus (MFG), the left Ventral Precentral Sulcus and the Left Supramarginal Gyrus for *PRO* compared to parasitic gaps. Fabre [25] further compared the contrast images between null pronouns, wh-traces and resumptive pronouns, and observed a shared network of temporal areas for establishing an intra-sentential dependency irrespective of its syntactic encoding.

Taken together, the fMRI literature has mainly associated the Angualr Gyrus, the MTGs, the STG/STSs, the ATLs, the IFGs and the Precentral Gyrus with pronoun resolution (See Figure 4.1 for a summary, taken from Fabre [25]). Therefore, we expect to replicate these finding in our naturalistic paradigm.

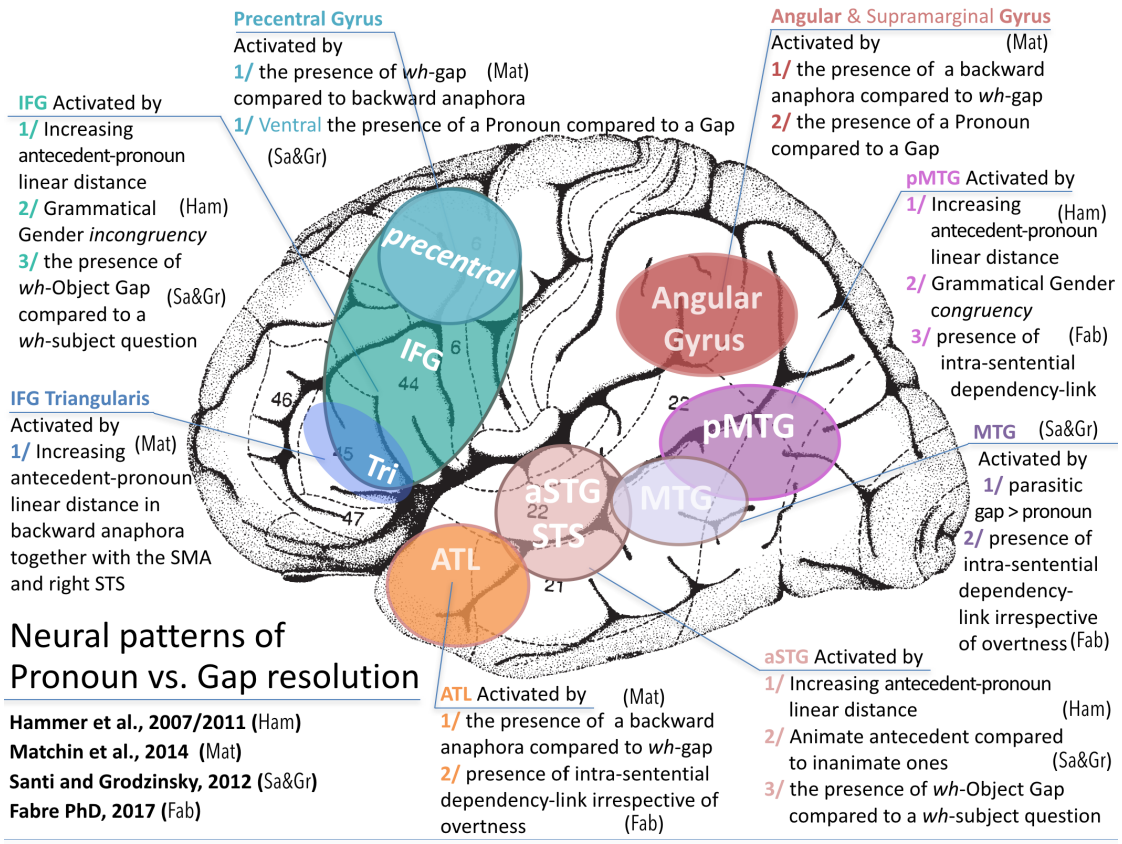


Figure 4.1: Summary of fMRI findings on pronoun resolution, from Fabre [25].

Part III

Computational Models

CHAPTER 5

APPROACHES TO PRONOUN RESOLUTION

Computational models for pronoun resolution can be mainly divided into theory-driven models and corpus-driven models. Most of the early approaches to pronoun resolution are based on theoretical proposals such as syntactic constraints and discourse salience as discussed in Chapter 2. With the availability of annotated coreference corpora in the mid-1990s, corpus-based models have become the current trend in the coreference resolution research. This chapter briefly reviews the most representative models from the theory-driven and the corpus-driven approaches and describes the rationale for the models that were selected for testing in the current study.

5.1 Theory-driven models

5.1.1 Syntax-based models

The earliest and best-known syntax-based algorithm on pronoun resolution is the Hobbs algorithm [49]. This algorithm traverses the parsed syntactic trees of the current and previous sentences in a left-to-right, breadth-first order and searches for an antecedent that is matched in gender and number. The Hobbs algorithm incorporates the locality constraints in the Binding Theory as discussed in Section 2.1, and gives preferences for grammatical subject of the sentence (see Chapter 6 for a detailed description of the Hobbs algorithm).

5.1.2 Salience-based models

One early influential pronoun resolution model based on the Centering Theory is proposed by Brennan et al. (1987; henceforth BFP). The BFP algorithm computes the preferred antecedents from relations that hold between the forward and backward looking centers in adjacent sentences. The algorithm first generates all possible $C_b - C_f$ pairs for the pronoun in Utterance U_n. It then filters all pairs based on the Centering rules, For example, C_b must be pronominalized if any C_f is pronominalized; C_b is the highest ranked elements in the list of C_fs, etc. Finally, the algorithm ranks the remaining pairs by transition orderings, where maintaining the same C_b (Continue) is preferred over maintaining the same C_b in U_{n+1} but not in U_{n+2} (Retain), which is preferred to changing C_b in U_{n+1} (Shift). The selected $C_b - C_f$ pair is the most preferred relation according to the transition order.

Another influential salience-based model for pronoun resolution is the RAP algorithm proposed by Lappin and Leass [63]. Unlike the BFP algorithm that compares a discrete number of centers, the RAP algorithm assumes a graded activation level for each entity in the discourse. It also follows a generate-filter-rank procedure and takes as input the output of a full parser and filters entities according to binding constraints and gender/number agreement. It then assigns a salience weight to each entity depending on its recency, syntactic position, grammatical role, etc. The initial weights for each factor that contributes to the salience of an entity are derived from extensive corpus experiments, as shown in Table 5.1. These weights are then halved for each sentence boundary in between the entity and the pronoun, and the weights for all occurrence of the same entity are summed. The entity that receives the highest salience weight is the antecedent

of the pronoun.

Factor type	Initial weight
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object and oblique complement emphasis	40
Head noun emphasis	80
Non-adverbial emphasis	50

Table 5.1: Initial weights for salience factors in the RAP algorithm [63]

Both the BFP and the RAP algorithms incorporate claims from the Centering Theory, and the RAP algorithm also includes the Binding Principles and gender/number agreements. A more recent salience-based model proposed by van Rij et al. [107] is built within the cognitive architecture of Adaptive Control of Thought-Rational (ACT-R) [3]. Similar to the RAP algorithm, the ACT-R model for pronoun resolution also assumes a graded salience for each entity and ranks the activation level for each mention, but the activation level is only based on recency, frequency and grammatical role of the mention. More frequent and more recent entities have higher base-activation level, and entities in a subject position have higher spread-activation level (see Chapter 7 for a detailed description of the ACT-R formula).

5.2 Corpus-driven models

5.2.1 Machine-learning models

The advent of the MUC-6 [37] and MUC-7 [15] coreference corpora encourages machine-learning models on pronoun resolution. The most influential paradigm in the early machine-learning approach is the mention-pair model for pronoun resolution [97]. This model first classifies every mention of the text into pairs of antecedent and then decides whether the pairs can be clustered based on transitivity. The model contains 12 features that represents discourse, morphological/lexical, semantic and syntactic information (see Table 5.2). This feature set has been considered the core feature set in subsequent coreference resolution systems. The learning algorithm used in Soon et al. [97] is C5, a commonly used decision tree algorithm. This algorithm considers mention j starting from the second one as a possible anaphor, and every mention i before j as a possible antecedent. For every mention pair i and j, a feature vector is generated and given to the decision tree classifier. The model takes the immediately preceding j and proceeds in the reverse order of the mentions in the document until the classifier returns true or there is no remaining mention to be tested. Soon et al. [97] evaluated their model on the MUC-6 and MUC-7 coreference corpora and achieved competitive accuracy compared to the rule-based algorithms.

To further explore the effect of feature set, Ng and Cardie [77] added 41 features based on common-sense knowledge and linguistic theories on top of Soon et al.'s 12 features. These features included more complex string matching features, finer-grained semantic compatibility features, more NP-type and grammatical role features, and more sophisticated syntactic constraints such

Feature Type	Description	Value
Discourse	Number of sentences in between mention i and j	Integer
Morphological/lexical	Do mentions i and j match in string?	Boolean
	Is mention i an alias of mention j or vice versa?	Boolean
	Is mention i a pronoun?	Boolean
	Is mention j a pronoun?	Boolean
	Is mention j a definite noun phrase?	Boolean
	Is mention j a demonstrative noun phrase?	Boolean
	Are mentions i and j both proper names?	Boolean
	Do mentions i and j agree in number?	Boolean
	Do mentions i and j agree in gender?	Boolean
Semantic	Do mentions i and j belong to the same semantic class?	Boolean
Syntactic	Is mention j in apposition to mention i?	Boolean

Table 5.2: Feature set in Soon et al.'s (2001) mention-pair model for pronoun resolution. Mention *i* is the candidate antecedent; mention *j* is the anaphor

as the Binding Theory. Ng and Cardie [77] evaluated their model on MUC-6 and MUC-7 and found a decrease in precision for the common nouns using the full feature set compared to Soon et al. [97]. They then dropped the features that led to low precision score for common nouns and manually selected 18 additional features. They retrained the classifier using the reduced feature set and the results showed significant increase in performance compared to Soon et al.'s model. However, there is a substantial drop of performance in precision for pronouns. This discrepancy seems to suggest that separate classifiers are needed for pronouns and common nouns.

Although the mention-pair model is the most influential learning-based coreference model, it has been criticized as only considering local information between two mentions. Since information extracted from two mentions and their local contexts may not be sufficient to determine coreference, especially if the antecedent is semantically empty (e.g., a pronoun) or lacks gender/number

information (e.g., *Clinton*), the mention-pair model may not perform well in merging mention pairs into clusters. For example, if there are three mentions in a document: *Mr. Clinton*, *Clinton* and *she*, the mention-pair model may determine that *Mr. Clinton* and *Clinton* are coreferential using string-matching features, and that *Clinton* and *she* are coreferential based on proximity and no evidence for gender/number disagreement. Then due to transitivity, the model will wrongly merge the two pairs into a cluster {*Mr. Clinton*, *Clinton*, *she*}, even though *Mr. Clinton* and *she* mismatch in gender. This weakness of mention-pair models motivated the entity-mention model, which determines whether the current mention belongs to a preceding coreference cluster. Take the *Clinton* example again, when encountering *she*, the model takes into consideration that *Mr. Clinton* and *Clinton* are already in the same cluster, and determines that *she* does not belong to the cluster because there is a gender mismatch between *she* and *Mr. Clinton* in the cluster. Therefore, the entity-mention model is able to enforce global coherence across mention pairs.

5.2.2 Neural coreference models

With the development of representing words as vectors that convey semantic dependencies [see e.g., 75], deep learning approaches to coreference resolution has been developed [18, 19, 64, 112]. Clark and Manning's [18] algorithm was based on entity-level information. Its architecture consisted of mainly four sub-parts: the *mention-pair encoder* passes features through a fully connected feed-forward neural network to produce distributed representations of mentions; the *cluster-pair encoder* uses pooling over mention pairs to produce distributed representations of cluster pairs; *mention-ranking model* scores the candidate antecedents

to feed the *cluster-ranking model*, which scores pairs of clusters by passing their representations through a single-layer neural network.

The features used for the model included the average embeddings of words in each mention, binned distance between the mentions, head word embedding, dependency parent, embeddings of the first, last and two preceding words of the mention, average embeddings of 5 preceding and succeeding words of the mention, type of mention, position of mention, length of the mention, document genre, string match, etc. (see Chapter 8 for a detailed description of the Clark and Manning neural network model). The Clark and Manning's neural network model was trained on the CoNLL-2012 Shared Task [85] and achieved an $F1$ score of 65.39 on the CoNLL English task and 63.66 on the Chinese task.

5.3 Summary of the models

All the rule-based approaches such as the Hobbs algorithm, the BFP algorithm and the RAP algorithm are heavily knowledge-based and focused only on pronouns. The Hobbs algorithm assumes perfect syntax knowledge and the BFP algorithm assumes discourse knowledge according to the Centering Theory; the RAP algorithm can be seen as a hybrid model as it was both syntax- and discourse-based. Therefore, both the Hobbs and the RAP algorithm are dependent on accurate parsing of the sentences. Another problem with the RAP algorithm is its weight assignment scheme for its salience factors. These weights are corpus-dependent, thus it remains a question whether they are still valid for pronouns resolution in another corpus or in another language.

The machine-learning and deep learning models, on the other hand, are

knowledge-poor algorithms as they aim to reduce the level of dependency on rules and external knowledge. These algorithms learn feature representations from corpora and are based on few hand-engineered features. The machine learning and deep learning approaches do not confined to pronouns only, instead, they predict coreference relations between all NPs. Since these approaches view coreference resolution as a cluster problem, there lacks a standard evaluation metric to compare the machine learning and deep learning models with the Hobbs algorithm, which do not cluster antecedents. In addition, compared to the rule-based algorithms, the machine-learning and deep learning models are more restricted on the genre of the corpora because they are trained on either the MUC or the CoNLL corpora, which mainly contain news articles. It is questionable whether these trained weights could apply to text of another genre, such as narratives or everyday conversations.

5.4 Models tested in the current study

The current study selects three computational models to test whether they could indicate cognitive states during pronoun resolution in the human brain. The Hobbs algorithm is selected to examine the cognitive reality of the syntactic constraints and morphological agreement; the ACT-R model tests how salience of the antecedent influences real time pronoun resolution. ACT-R is selected over the other salience-based models as it is adapted from the cognitive architecture of ACT-R, which is specifically intended for human cognition. In addition, the ACT-R model is not dependent on syntactic structures as does the RAP algorithm, hence it is not influenced by accuracy of the parser.

Finally, we selected the Clark and Manning [18] neural network model as a comparison to the two theory-driven models to examine whether pronoun resolution could be viewed as an emergent property from the brain's responses to the task it is presented. This neural network model is selected over Wiseman et al.'s [112] and Lee et al.'s [64] model because it has been trained on both the English and Chinese data in the CoNLL corpus and is ideal for the comparison of English and Chinese pronoun resolution in the human brain. The following Chapters 6 to 8 discusses the three computational models in detail. Section 9.1 provides a detailed summary of all the elements involved in the three selected models.

CHAPTER 6

THE HOBBS ALGORITHM

6.1 The Algorithm

The Hobbs algorithm, originally presented in Hobbs [49], depends only on a syntactic parser plus a morphological gender and number checker. The input to the Hobbs algorithm includes the target pronoun and the parsed trees for the current and previous sentences. The algorithm searches for a gender and number matching antecedent by traversing the trees in a left-to-right, breadth-first order, that is, it starts at the tree root and explores the neighboring nodes at the present depth prior to moving on to the nodes at the next depth level. If no candidate antecedent is found in the current tree, the algorithm searches on the preceding sentence in the same order. The steps of the Hobbs algorithm are as follows:

1. Begin at the NP node immediately dominating the pronoun.
2. Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.
3. Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node that is encountered which has an NP or S node between it and X.
4. If node X is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as antecedent. If X is not the highest S node in the sentence, continue to step 5.

(5) From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.

(6) If X is an NP node and if the path p to X did not pass through the $\bar{N}$ node that X immediately dominates, propose X as the antecedent.

(7) Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.

(8) If X is an S node, traverse all branches of node X to the right of path p in a left-to-right. breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.

(9) Go to step 4.

Figure 6.1 illustrates the search algorithm. There are two pronouns *they* in the sentence taken from "The Little Prince". To search for the antecedent for the second *they* under NP8, the algorithm first goes to node NP8 that directly dominates the pronoun following Step (1) of the algorithm, it then goes to node S9 following Step (2) of the algorithm. From S9, the algorithm skips Step (3) and (4) and goes to node S8, and from S8 to to S4 following Step (5). Then according to Step (3), the algorithm proposes the first encountered NP7, namely, the first pronoun*they* as the antecedent.

The Hobbs algorithm conforms to the Binding Theory as it always searches the antecedent in the left of the NP (Principle B: Step (3)) and do not go below any NP or S node encountered (Principle A: Step (8)). It also respects gender, person, and number agreement and captures recency and grammatical role preferences in the order it performs the search as the leftmost NP is usually the subject of a sentence. Hobbs [49] evaluated his algorithm on 300 examples containing third person pronouns, and it worked in 88.3% of the cases. With some selectional

constraints on dates and location antecedents (i.e., restricting dates and location NPs such as *2018* and *school* to be the antecedent of *it*), the algorithm achieved an accuracy of 91.7%.

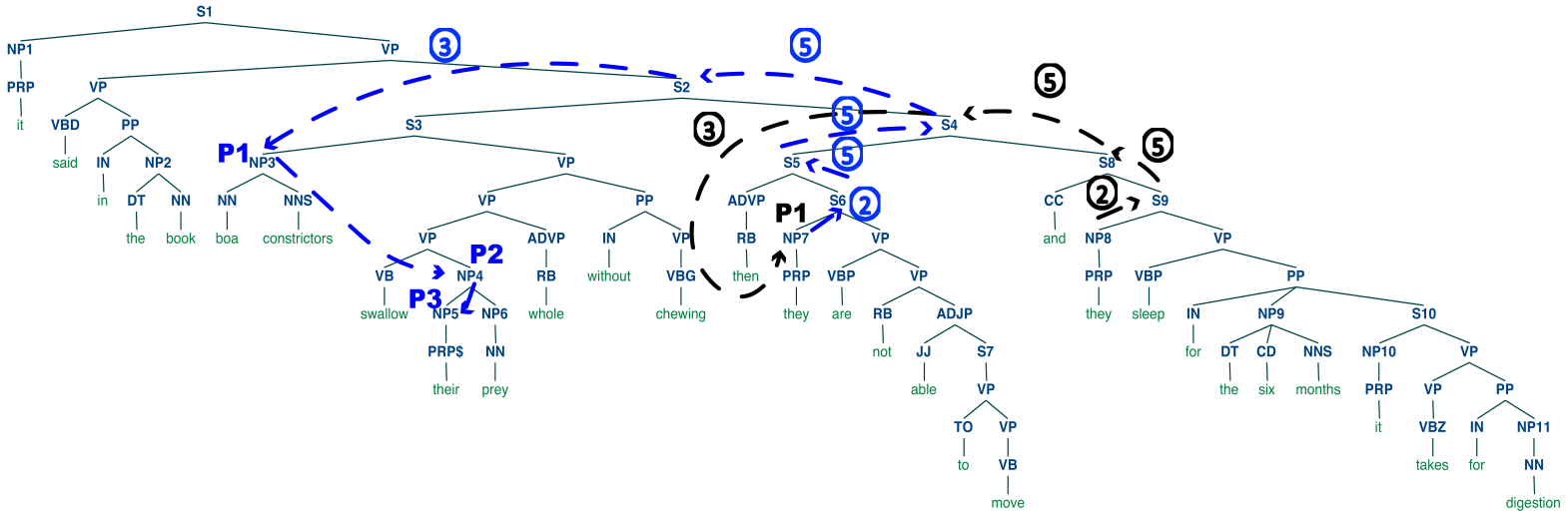


Figure 6.1: Illustration of the Hobbs algorithm in English. The sentence is taken from *The Little Prince*, which contains two pronouns "they". The black dashed line points to the search steps for the second "they" under NP8, and the blue dashed line demonstrates the search procedure for the first "they" under NP7 following a similar procedure. The circled number indicates the Steps in the algorithm. "P1-3" indicates the number of proposals, i.e., the Hobbs distance.

6.2 Hobbs distance

A major problem with the Hobbs algorithm is when there are competing antecedents. As shown in Figure 6.1, for the first pronoun *they* under NP7, NP3 *boa constrictors*, NP4 *their prey* and the possessive *their* under NP5 are all potential antecedents. In this case, NP3 and NP5 refer to the same entity and form a referential chain {*boa constrictors*, *their*, *they*, *they*}. But NP4 *their prey* is not in the referential chain. However, the Hobbs algorithm will always choose the left-most NP if it agrees with the pronoun in gender and number, in this case, NP3. This leads to wrong prediction if NP3 is not in the referential chain, or if we want the algorithm to always predict the immediate preceding antecedent (NP5).

To accommodate the problem with competing antecedents, the notion of "Hobbs distance" has been proposed. Hobbs distance refers to the number of proposals that the Hobbs algorithm has to skip, starting backwards from the pronoun, before the potential antecedent NP is found [see 58, p.721]. For example, in Figure 6.1, the algorithm first proposes NP3 *boa constrictor* as the antecedent for *they* under NP7. This proposal (P1), though correct, is not the immediately preceding antecedent, which should be the possessive pronoun *their* under NP5. So if we set the algorithm to always predict the immediate preceding antecedent, it would have to keep running until it reaches NP5, which is the third proposal (P3). The second proposal is NP4 which is at a higher level than NP5. The number of proposals that the algorithm skips until it reaches the immediate antecedent is 3, so the Hobbs distance between the immediate antecedent *their* and the pronoun *they* under NP7 is 3. Similarly, the Hobbs distance for the second *they* under NP8 is 1 since the first proposal is the correct immediate antecedent.

The Hobbs distance metric has been integrated in Ge et al. [32]'s statistical model for pronoun resolution to calculate the probability of candidate antecedents. Other factors in the model include gender/number/animaticity of the candidate antecedent, governing head information and noun phrase repetition. The Ge et al.'s model has been tested on 21 million words of Wall Street Journal text and achieved an accuracy of 84.2%. This test dataset is much larger than that in the original Hobbs's paper, which contains only 300 hand selected sentences from a novel, an essay and news articles.

6.3 Hobbs algorithm applies to Chinese

The Hobbs algorithm, when applied to third person pronouns in spoken Chinese, no longer contains a gender and number agreement checker because Chinese pronouns (spoken form) do not distinguish gender and Chinese NPs do not mark plurals (see Chapter 3). It is therefore expected that the performance of the algorithm would degrade for Chinese pronoun resolution. For the Hobbs distance metric, it is hypothesized that the overall Hobbs distance for the correct immediate antecedent in Chinese would be higher than that in English, because the number of competing NPs increases due to the lack of a gender/number agreement filter.

In Chapter 9 we presented a comparison of the Hobbs algorithm performance for third person pronoun resolution in the English and Chinese *The Little Prince*. Our results confirmed the hypothesis that the Hobbs algorithm performs worse in Chinese and the overall Hobbs distance between the correct immediate antecedent and the pronoun is higher in Chinese.

CHAPTER 7

THE ACT-R MODEL

7.1 ACT-R as a cognitive architecture

The term "cognitive architecture" refers to "a basic set of primitives out of which cognitive models may be built" [41]. Cognitive architecture of the thinking human is analogous to computer architecture which specifies the number and type of memories, instruction set, etc. The set of primitives in the cognitive architecture could be re-used to explain human performance on different tasks that calls upon different aspects of their intelligence. Many instances of cognitive architecture has been proposed in the literature, one famous example being Newell's (1990) Soar system, which continues to evolve and contributes to the cognitive science community since his death [see Chapter 5 of 41, for an introduction of cognitive architecture].

ACT-R [3, 4] is another cognitive architecture that aspires to connect human mind and brain. The full architecture of ACT-R consists of eight independent modules. The visual and the aural modules are the two perceptual modules that hold the representation of a problem, such as the representation of an equation "$3x - 5 = 7$"; the imaginal module holds a current mental representation of the problem. For instance, the intermediate representation of "$3x = 12$" in solving "$3x - 5 = 7$". The goal module keeps track of the current intentions in solving the problem, such as performing algebra transformation; and the declarative module retrieves critical information from declarative memory, such as "$7 + 5 = 12$". These three modules are connected via the central procedural module that recognizes and puts information in the buffers associated with the

them. The final two modules, the manual module and the vocal module are the response modules that program the output, such as "$x = 4$" . Figure 7.1 illustrates the interconnections among the modules in ACT-R [see Chapter 1 and 2 of 4, for an overview of the ACT-R modules].

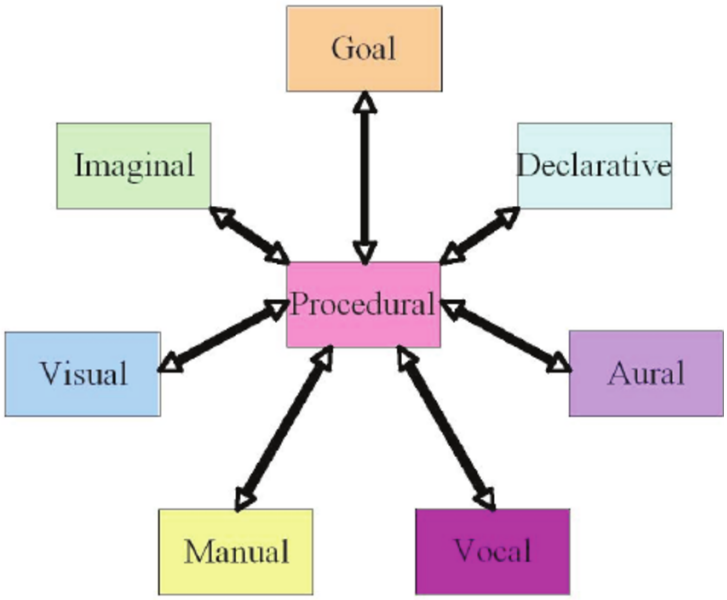


Figure 7.1: The interconnections among modules in ACT-R. From Anderson [4].

To connect ACT-R with states of computation in the human brain at the implementation level in the sense of Marr [69], Anderson [3] reported brain activity from an fMRI study where children of 11-14 years old learned to solve simple linear algebra equation involving zero, one, or two steps operations (e.g., $x = 4$, $3x = 12$, $3x - 5 = 7$) in the scanner over 5 days. Brain regions associated with a specific module of the ACT-R theory indicates the location of that module in the brain. Anderson [3] identified 5 brain regions roughly correspond to the 5 modules: the parietal region for the imaginal modules, the anterior cingulate region for the goal module, the prefrontal region for the declarative memory module, the caudate region for the procedural module and the the motor region for the manual module. Figure 7.2 illustrates the

correlation between the predicted brain activity and the actual BOLD response in the corresponding regions.

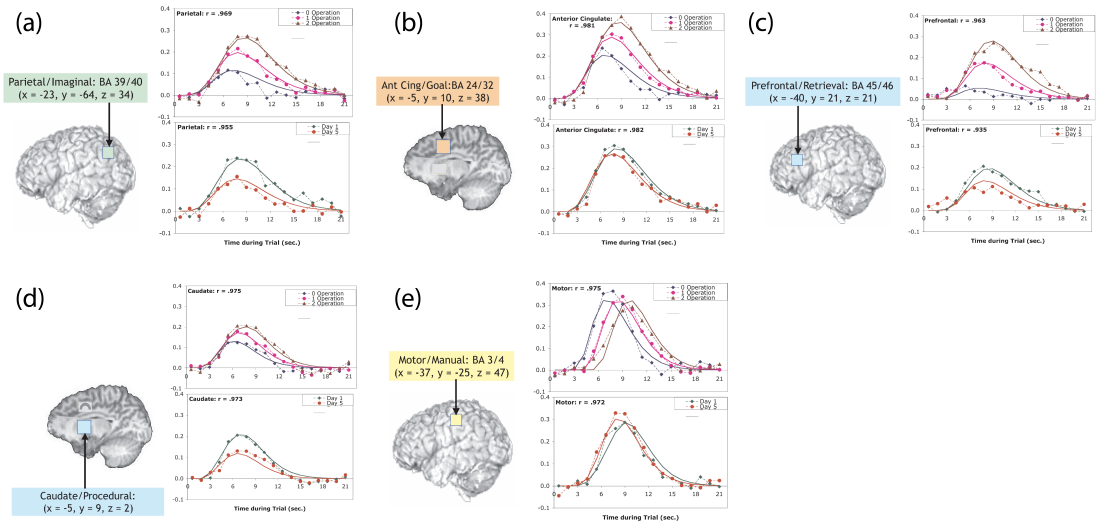


Figure 7.2: Module behavior associated with BOLD response in different brain regions: (a) the the imaginal modules predicts the parietal region; (b) the goal module predicts the anterior cingulate region; (c) the declarative memory module predicts the prefrontal region; (d) the procedural module predicts the caudate region, and (e) the manual module predicts the motor region. The solid lines represents predicted BOLD response by the ACT-R theory and the dashed lines are the actual BOLD signals. From Anderson [3].

The ACT-R architecture has been applied to explain the moment-by-moment working-memory retrievals and associated control structure that subserves sentence comprehension by Lewis and Vasishth [65]. They argue that many difficulty effects in sentence processing, such as the garden-path effects, can be attributed to memory decay and interference effect described in the declarative module of the ACT-R theory. Section 7.2 reviews in detail the declarative memory module in ACT-R, which is mostly relevant to complexity in sentence processing and is adapted by van Rij et al.'s (2013) ACT-R model for pronoun resolution (see Section 7.3).

7.2 The declarative module in ACT-R

In the modular system of ACT-R, declarative memories of past events are stored as "chunks" in the buffer of the declarative module. For example, in the linear algebra equation case, the relevant memory chunk could be "7 + 5 = 12". The chunks have activation levels that determine the speed and success of their retrieval, and the activation level reflects both the inherent strength of the memory and the strength of its association to elements in the current context. The formula to calculate the activation level of chunk i is given below:

$$A_i = log(\sum_{k=1}^{n} {t_k}^{-d}) + \sum_{j=1}^{m} W_j \times S_{ji}$$

The first part of the equation $log(\sum_{k=1}^{n} {t_k}^{-d})$ computes the inherent strength, or the base-level activation of chunk i, which reflects the past history of usage of chunk i. t_k is the time passed since the kth representation of i, and each representation decays over time as a negative power function ${t_k}^{-d}$. The parameter d is set to 0.5 as the default value in ACT-R based on a range of experiments to model human performance in memory retrieval tasks. Different representations of the same chunk i adds up to reflect the effect of practice.

The second part of the equation $\sum_{j=1}^{m} W_j \times S_{ji}$ reflects the associative activation that chunk i receives from the context elements j. S_{ji} is the strength of association between j and i, namely, how much the presence of elements j makes chunk i more probable. Currently, the default value of S_{ji} in ACT-R is set to 2. W_j is the attentional weighting which equals to W/n where n is the number of sources of activation. This attentional weighting equation sets the sum of attentional weights to 1 [see Chapter 3 of 4, for a detailed explanation of the activation equation in ACT-R].

The activation equation corresponds to two statistical effects of memory retrieval: (1) The more often and more recent a memory occurs, the more likely it is to be retrieved in the future. This reflects the practice effect in the base-level activation part of the equation. (2) The more memories associated with a particular cue, the worse it is to predict a particular memory. This is reflected in the associative activation of the equation.

The activation equation for memory retrieval has been tested in a series of sentence recall experiments [e.g., 5, 84], and the activation level fits well with latency in the recall task over time. Lower activation level of a memory in the recall experiment has also been shown to be associated with higher BOLD responses in the left prefrontal region [96], consistent with Wang et al.'s (1992) finding that the left prefrontal region is predictive of memory for words.

7.3 The ACT-R model for pronoun resolution

The declarative memory module of ACT-R directly relates to pronoun resolution as the mapping between the pronoun and its antecedent can be viewed as memory retrieval modulated by the activation levels of the antecedent. Using the same primitives of the memory module in ACT-R, van Rij et al. [107] proposed an ACT-R model for pronoun resolution which reflects three factors that influences successful retrieval of the antecedents – frequency, recency, and the grammatical role of the antecedent. The formula for the activation level for the antecedent i of a pronoun is exactly the same with the activation equation for memory retrieval in ACT-R:

$$A_i = log(\sum_{k=1}^{n} t_k{}^{-0.5}) + \sum_{j=1}^{m} 1/n \times 2$$

The base-level activation $log(\sum_{k=1}^{n} t_k^{-0.5})$ represents frequency and recency of each mention of the antecedent i, and the associative activation $\sum_{j}^{m} W_j \times 2$ represents the influence of grammatical role of each mention. If mention j is a subject, it has a attentional weighting (W) of 1; which is divided by the total number of mentions of this antecedent n ($W_j = W/n$), as the total value of associative activation cannot be infinite. W_j is then multiplied by 2, the default value of associative strength S_{ji} in ACT-R.

The effects of frequency and recency are folded into the calculation of the base activation for antecedent i, such that the more mentions it has, and the more recent the mentions occur, the higher the base activation. Conversely, if antecedent i has been mentioned only once, or if its last mention was a long time ago, its activation level will be low, and it will rank lower on the activation list for all the candidate antecedents. Subjecthood of the mentions of antecedent i gains an associative activation in addition to the base activation. Overall, the amount of activation value of an entity in the discourse context is computed based on recency, frequency and grammatical role of the entity, and the highest ranked entity is predicted to be the antecedent of the pronoun.

To give a concrete example of how the activation level for each antecedent is calculated, in the English sentences "It said in the book boa constrictors swallow their prey whole without chewing, then they$_1$ are not able to move and they$_2$ sleep for the six months it takes for digestion." (see Figure 6.1), the previous mentions of the pronoun *they*$_2$ are *they*$_1$, *their* and *boa constrictors*. The time elapsed from these three previous mentions to *they*$_2$ in the audio are 5.42 s, 4.76 s and 1.74 s respectively. Since *boa constrictor* is a subject of a subordinate clause,

it gets an associative weighting W of 1. Therefore, the activation level for the antecedent of $they_2$ is calculated as:

$$\begin{aligned} A_{they_{15}} &= log(5.42^{-0.5} + 4.76^{-0.5} + 1.74^{-0.5}) + 0/3 \times 2 + 1/3 \times 2 + 0/3 \times 2 \\ &\approx 1.17 \end{aligned}$$

Similarly, for the corresponding Chinese sentence "这本书中写道这些蟒蛇把它们$_1$的猎获物不加咀嚼地囫囵吞下, 尔后就不能再动弹了, 它们$_2$就在长长的六个月的睡眠中消化这些食物.", the previous mentions of the last pronoun 它们$_2$ is 它们$_1$ and 蟒蛇. The time elapsed from them to 它们$_2$ in the audio are 5.37 s and 0.44 s, respectively. The mention 蟒蛇 is in a subject position so it gets an attentional weighting of 1. The activation level for the antecedent of the pronoun 它们$_2$ is therefore calculated as:

$$\begin{aligned} A_{它们_{11}} &= ln(5.37^{-0.5} + 0.44^{-0.5}) + 1/2 \times 2 + 0/2 \times 2 \\ &\approx 1.66 \end{aligned}$$

The three elements in the ACT-R module is consistent with the salience-based account on pronoun resolution such as the Centering Theory [39] and the Accessibility theory [7]. The fit of the model prediction with human self-paced reading data in van Rij et al. [107] also supports memory retrieval as a key cognitive mechanism in pronoun resolution. The ACT-R model for pronoun resolution assumes no knowledge of syntactic structure and morphological agreement. The only linguistic knowledge needed in the ACT-R model is the notion of grammatical role. This makes the ACT-R model for pronoun resolution more universal cross-linguistically than the Hobbs algorithm and can be applied to pronoun resolution in Chinese without language-specific modification. Chapter 9 evaluates

the performance of the ACT-R model for third person pronoun resolution in the English and Chinese translation of the book *The Little Prince*.

CHAPTER 8

THE NEURAL COREFERENCE MODEL

8.1 The architecture of the neural coreference model

The neural network model for coreference resolution [18, 19] deals with both pronominal and nominal coreference relations. The model consists of a "mention-pair encoder", a "cluster-pair encoder", a "mention-ranking model" and a "cluster-ranking model". The mention-pair encoder generates distributed representations for pronoun-antecedent pairs, or mention pairs, by passing relevant features through a three-layer feed-forward neural network. The cluster-pair encoder generates distributed representations for pairs of clusters through a pooling operation over representations of relevant mention pairs. The mention-ranking model scores the candidate antecedents to prune the set of possible antecedent and the cluster-ranking model scores coreference compatibility for each pair of clusters.

The input layer of the neural network model consists of a large set of features including word embeddings for the mention pairs, type and length of the mentions, linear distance between the mention pairs, etc. (see Table 8.1). These feature vectors are concatenated to produce an I-dimensional vector $h_0(a, m)$ as the representation for the mention m and the antecedent a.

Feature Type	Description
Word embedding	head word
	dependency parent
	first word
	last word
	two preceding words
	two following words
	averaged of the five preceding words
	averaged of five following words
	all words in the mention
	all words in the mention's sentence
	and all words in the mention's document
Mention	type (pronoun/noun/proper name/list)
	position in the document
	contained in another mention or not
	length of the mention in words
Document	genre (broadcast news/newswire/web data)
Distance	intervening sentences
	number intervening mentions
	mentions overlap or not
String matching	head match
	exact string match
	partial string match

Table 8.1: Feature set of the neural network model for coreference resolution. From Clark and Manning [18].

The input layer then passes through three hidden layers of rectified linear units (ReLU), and the output of the last hidden layer is the vector representation for the mention pair $r_m(a, m)$.

$$h_i(a, m) = ReLU(W_i h_{i-1}(a, m) + b_i)$$

For pairs of clusters $c_i = \{m_1^i, m_2^i, ..., m_{c_i}^i\}$ and $c_j = \{m_1^j, m_2^j, ..., m_{c_j}^j\}$, the cluster-pair encoder first forms a matrix $R_m(c_i, c_j) = [r_m(m_1^i, m_1^j), r_m(m_2^i, m_2^j), ..., r_m(m_{c_i}^i, m_{c_j}^j)]$, then applies a pooling operation over $R_m(c_i, c_j)$ to produce a distributed representation for the cluster pair $r_c(c_i, c_j)$. The mention-ranking model assigns a score for each mention pair by applying a single fully connected layer of size one on the mention pair representation $r_m(a, m)$. The model is then trained with the max-margin training objective.

$$s_m(a, m) = W_m r_m(a, m) + b_m$$

Similarly, the cluster-ranking model assigns a coreference score for each cluster pair and an anaphoricity score for mention m (i.e., how likely mention m has an antecedent). These scores are used to decide whether mention m should be merged with one preceding cluster or not during testing. Figure 8.1 illustrates the architecture of the mention-pair and the cluster-pair encoder in Clark and Manning [18].

$$s_c(c_i, c_j) = W_c r_c(c_i, c_j) + b_c$$

$$s_{NA}(m) = W_{NA} r_m(NA, m) + b_{NA}$$

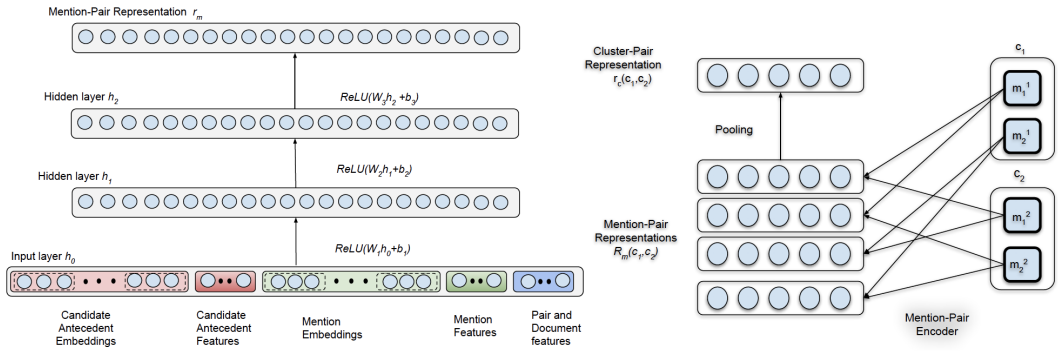


Figure 8.1: The mention-pair and the cluster-pair encoder in Clark and Manning's (2016a) neural coreference system.

8.2 Performance of the model

The neural network model encodes no syntactic tree structures, but it captures some semantic information in its word embedding features. It also contains some discourse-level information such as linear distance between the mention pairs across several sentences. Clark and Manning [18] trained the model on the CoNLL-2012 Shared Task [85] and it achieved state-of-the-art results with an F1 score of 65.39 for the CoNLL English task and an F1 score of 63.66 for the Chinese task.

The neural network model was evaluated on both pronominal and nominal coreference resolution, however, pronouns and full noun phrases (NPs) may rely on different sets of features. For example, string matching and measures for semantic similarity are powerful features for nominal coreference resolution, but are not applicable for pronoun resolution as word embeddings do not represent pronouns well. Theoretically, pronouns may serve a different discourse function from that of full NPs as full NPs introduce new entities in the discourse and

pronouns maintain the reference [91]. Based on these arguments, it is reasonable to expect different performance of the model on pronoun resolution and full NP coreference resolution.

In addition, as discussed in Chapter 5, the performance of the neural coreference model may be influenced by the genre of text as it is trained a corpus which mainly contains news articles. Chapter 9 goes on to evaluate the model on English and Chinese translations of "The Little Prince".

CHAPTER 9

MODEL COMPARISON

This chapter evaluates the performance of the Hobbs, the ACT-R and the neural coreference model for third pronoun resolution on English and Chinese translations of the book "The Little Prince". We then present an error analysis which details the relative strength and weakness of each model applied to the English and Chinese data.

9.1 Elements in the Hobbs, ACT-R and neural coreference models

As described in Chapters 6 to 8, the Hobbs algorithm, the ACT-R model and the neural coreference model focus on different aspects of the process of pronoun resolution. The Hobbs algorithm is syntax-sensitive; it relies on both parsed syntactic trees and morphological matching. The ACT-R model is based on principles of memory retrieval modulated by salience of the antecedent. The elements in the ACT-R model include recency, frequency and grammatical role of the past mentions of the antecedent. Both the Hobbs and the ACT-R model are relatively simple models that focus on one aspect of pronoun resolution and contain relatively few elements. In contrast, the neural coreference model is a complex system that incorporates a large set of features, ranging from lexical semantics and string matching to discourse-level information such as linear distance between the antecedent and the pronoun. The discourse genre and the speaker identity features in the original model [18, 19]are not used in this study, since they make no useful distinctions within "The Little Prince".

Table 9.1 lists the different components in the three computational models on pronoun resolution. Elements addressing same type of features, such as average word embeddings of the preceding and succeeding two words and average word embeddings of the preceding and succeeding five words in the neural coreference model are grouped together in this table.

Hobbs algorithm	ACT-R model	Neural coreference model
syntactic structure	frequency of antecedents	average embeddings of words in each mention
gender agreement checker	recency of antecedents	average embeddings of context words
number agreement checker	grammatical role of antecedents	position of mention in the discourse
		distance between mentions
		length of mention
		string matching between mentions

Table 9.1: Elements in the Hobbs, ACT-R and neural coreference model on pronoun resolution. Note that the gender and number agreement feature in the Hobbs algorithm does not apply to Chinese third person pronoun resolution.

9.2 Model performance on *The Little Prince*

9.2.1 The data

To evaluate the performance of the three models on third person pronoun resolution in the English and Chinese translation of *The Little Prince*, we first manually identified each mention (i.e., NPs) in the text and linked them with their coreferential mentions. The annotation is done using the annotation tool brat [99]. Figure 9.1 demonstrates sample annotations of a same sentence in the English and Chinese text.

Within the English audiobook text, 1755 pronouns (excluding possessives, reflexives, cleft and extraposition "it", and pleonastic "it") and 3127 non-

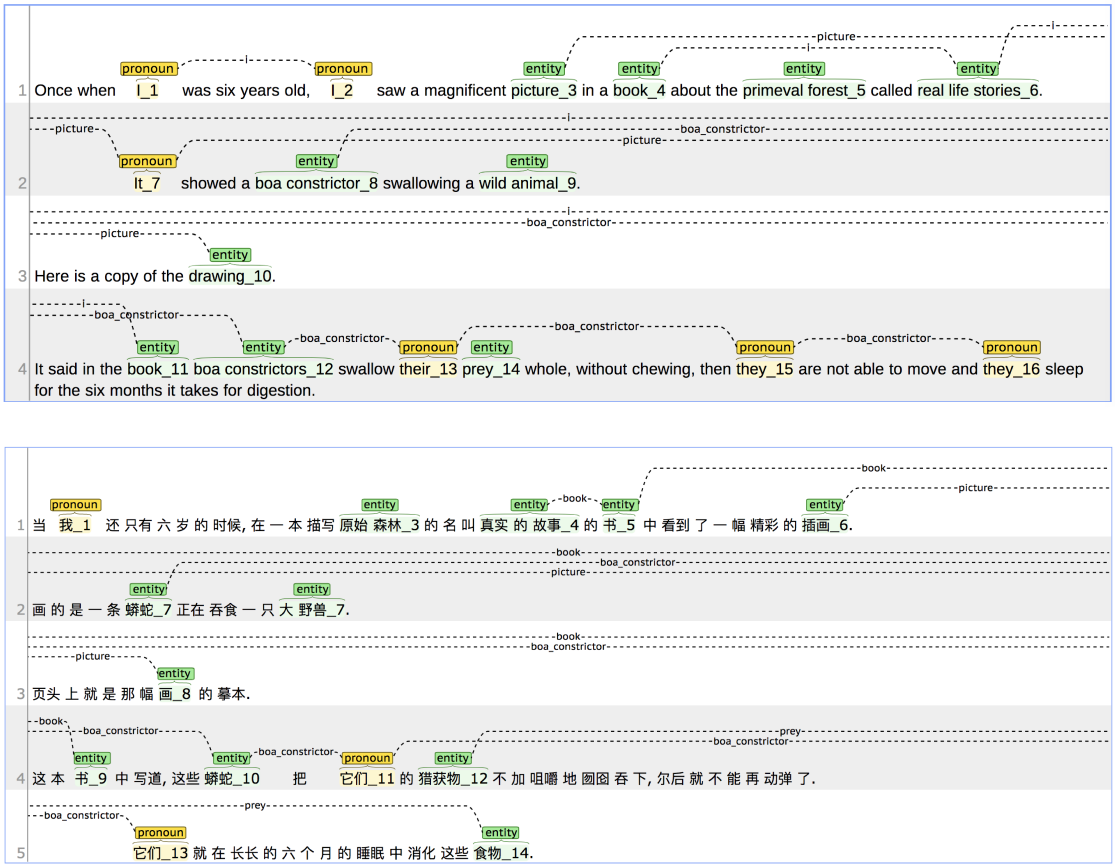


Figure 9.1: Sample annotations of pronouns and non-pronoun mentions in *The Little Prince* in English and Chinese, visualized using the annotation tool brat [99].

pronominal entities (4882 mentions in total) are identified. Pronouns with sentential antecedents are removed. For example, in the conversation "That is funny where you live a day only last a minute." "It is not funny at all." *it* in the second sentence is removed from our pronoun set as it refers to the whole sentence "where you live a day only last a minute". We also excluded third person pronouns whose antecedents are first or second person pronouns. Most of these cases are direct speech from characters as in (17a). We focused on third person pronouns because they provide gender information in English but not in Chinese. In addition, third person pronouns have been suggested to differ from first and second person pronouns in that first and second person pronouns

mark proximity in space and third person pronouns are further away [6]. The resulting English dataset contains 645 first person pronouns, 302 second person pronouns and 446 third person pronouns (see Table 9.2 and Table 9.3).

The Chinese audiobook text contains 1785 pronouns (excluding possessives and reflexives) and 2947 non-pronominal mentions (4732 mentions in total). We further pruned the pronoun set to exclude pronouns with sentential antecedents, and third person pronouns whose antecedents are first or second person pronouns such as in (17b). The resulting Chinese pronoun set contains 639 first person pronouns, 298 second person pronouns and 388 third person pronouns (see Table 9.2 and Table 9.3).

(17) a. "I$_i$ should not have listened to her$_j$." He$_i$ told me$_k$ one day.

b. "wo$_i$ bu gai tingxin ta$_k$ de hua." You yi tian, ta$_i$ gaosu wo$_j$ shuo.

"I$_i$ not should listen her$_k$ DE words." One day, he$_i$ told me$_j$.

	English		Chinese		
1st	i	me	我(wo)		
	505	121	621		
	we	us	我们(women)		
	16	3	18		
2nd	you		你(ni)		
	302		261		
			你们(nimen)		
			37		
3rd	she	her	她(ta)		
	41	14	62		
	he	him	他(ta)		
	268	64	303		
	it		它(ta)		
	136		73		
	they	them	她们(tamen)	他们(tamen)	它们(tamen)
	94	58	2	74	15

Table 9.2: Attestations of each pronoun type in the English and Chinese texts. Note that Chinese third person pronouns are homophones.

	English	Chinese
1st	645	639
2nd	302	298
3rd	446	388

Table 9.3: Attestations of each pronoun type in the English and Chinese texts after the pruning criteria.

As can been seen in Table 9.2 and Table 9.3, the number of first and second person pronouns are comparable in the English and Chinese texts, however, there are much fewer third person neutral pronoun *ta(it)* in the Chinese text. A closer scrutiny of the data revealed that Chinese tends to avoid using *ta(it)* to

refer to inanimate entities. For example, *my drawing* in the English sentence (18a) is a relative clause *what I drew* in the Chinese sentence (18b), and it is referred to by a zero pronoun *pro* in (18b).

(18) a. My drawing$_i$ was not of a hat. It$_i$ showed a boa constrictor digesting an elephant.

b. Wo hua de bu shi maozi. Shi yi tou jumang zai xiaohua yi tou daxiang.

I draw DE not is hat. *pro* Is one CL boa constrictor ASP_progressive digest one CL elephant.

9.2.2 Evaluation metric

To evaluate the performance of the three computational models, we first computed the Hobbs distance for each of the third person pronouns. A Hobbs distance of 1 indicates correct prediction of the Hobbs algorithm. For the ACT-R and the neural coreference model, we first calculated the activation levels and the neural coreference scores for all the preceding mentions for each third person pronoun. We used the pre-trained weights in Clark and Manning [18] to generate the neural coreference score.

We then ranked the potential mentions according to their ACT-R activation levels or their coreference scores. The ACT-R and the neural coreference model are considered correct if the true antecedent is ranked within the top 3 of the list. This is the SUCCESS@N metric ($N = \{1, 2, 3\}$), first proposed by Kolhatkar and Hirst [e.g., 59] and also used in Marasović et al. [68]. SUCCESS@N is the proportion of instances where the gold answer—the unit label—occurs within a

system's first N choices. SUCCESS@1 is standard accuracy. The SUCCESS@N metric allows some degree of ambiguity in selecting the the referents, which parallels human performance during pronoun resolution.

9.2.3 Performance

Figure 9.2 shows the distribution of the Hobbs distance, the ACT-R activation level and the neural coreference scores for third person pronouns in "The Little Prince" in English and Chinese. An independent-samples t-test was conducted to compare the Hobbs distance between the correct antecedent and the third person pronouns in the English and Chinese texts. There was a significant difference in the Hobbs distance in English ($M = 1.59, SD = 2.88$) and Chinese ($M = 2.9, SD = 3.6$); $t(832) = -5.87, p < 0.001$. However, there is no significant difference between ACT-R activation level for the correct antecedents for third person pronouns in English ($M = 2.52, SD = 1.15$) and Chinese ($M = 2.59, SD = 1.09$), $t(832) = -0.88, p = 0.38$. The mean neural coreference scores for correct antecedents in the English ($M = 5.44, SD = 4.36$) and Chinese ($M = 5.34, SD = 4.4$) text are not significantly different either ($t(832) = 0.34, p = 0.74$). Pearson's r test revealed no significant correlation among the three metrics in English and Chinese (see Figure 9.3).

The accuracy for the Hobbs algorithm (i.e., Hobbs distance=1), the ACT-R and the neural coreference model based on SUCCESS@N ($N = \{1, 2, 3\}$) is given in Table 9.4. We can see that for the English data, the Hobbs algorithm performs better than the ACT-R and the neural coreference model. At the lower threshold of S@3, the Hobbs algorithm achieved an accuracy of 97%, while the ACT-R

model is only 64% accurate. The neural coreference model only achieved an accuracy of 42% at S@3. For the Chinese data, the Hobbs algorithm is also the best performing algorithm, achieving an accuracy of 76% at S@3, which is slightly better than the ACT-R model with an accuracy of 74%. The neural coreference model is only accurate on 38% of the cases.

Cross-linguistically, the Hobbs algorithm has higher accuracy for the English data than for the Chinese data, while the ACT-R model performs better on the Chinese data than on the English data. This result, together with the significant difference in the mean Hobbs distance, supports our hypothesis that English speakers are more sensitive to syntactic and morphological cues, whereas Chinese speakers rely more on discourse-level features such as salience of the entities.

The neural coreference model does not perform well for either the English or the Chinese text, although the accuracy is slightly higher for the English data. Compared with the high F1 score (0.74) for pronoun and full NP coreference resolution for the CoNLL-2012 English test data, this low accuracy suggests that pronominal and nominal coreference resolution may rely on different sets of features. String matching and semantic similarity, for example, may be less powerful for pronominal resolution. The bad performance of the neural coreference model may also be due to the genre of the data. We used the pre-trained weights from the original Clark and Manning [18] model, which is trained on news articles. This weight may not generalize well to text of a different genre such as a novel.

	Model	Accuracy (%)		
		S@1	S@2	S@3
English	Hobbs	70	92	97
	ACT-R	32	46	64
	Neural Coreference	17	30	42
Chinese	Hobbs	40	64	76
	ACT-R	35	51	74
	Neural Coreference	13	28	38

Table 9.4: Accuracy of the Hobbs, ACT-R and Neural Coreference model in *The Little Prince* based on the SUCCESS@N ($N = \{1, 2, 3\}$) metric.

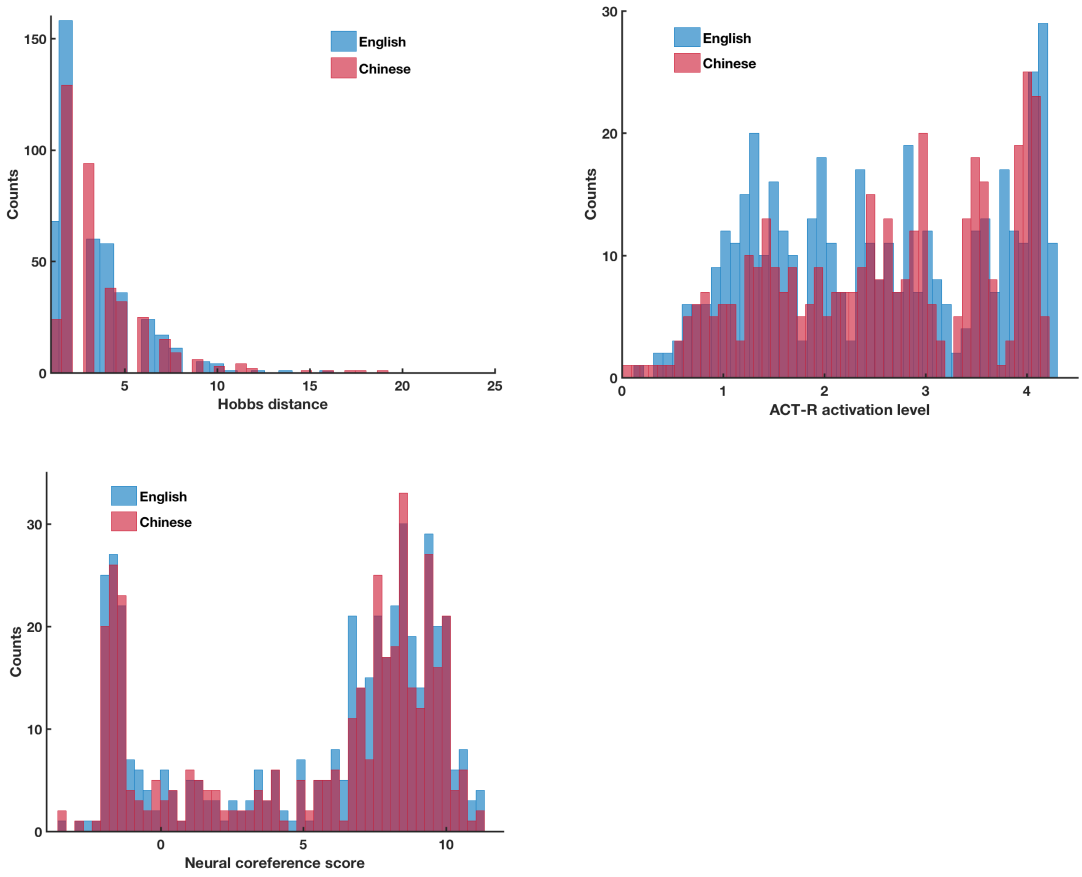


Figure 9.2: Histograms showing the distribution of the Hobbs distance, the ACT-R activation level and the neural coreference scores for third person pronouns in *The Little Prince* in English and Chinese.

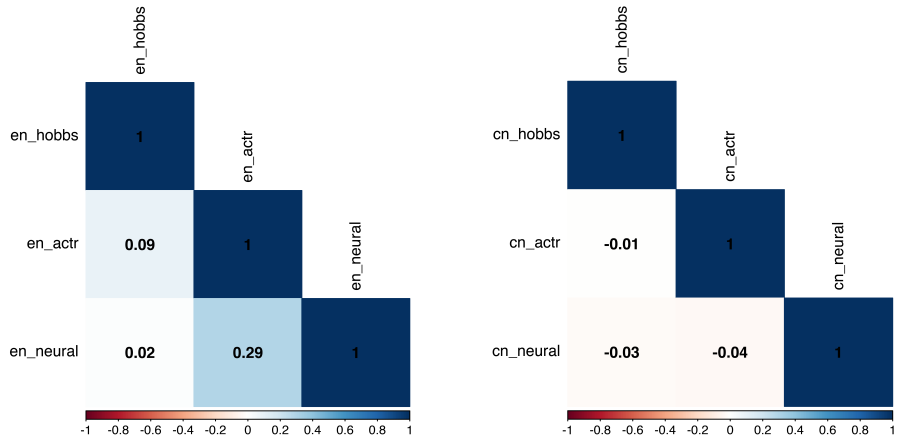


Figure 9.3: Correlation matrix of the Hobbs distance, the ACT-R activation level and the neural network coreference scores for third person pronouns in *The Little Prince* in English and Chinese.

Part IV

fMRI Experiment

CHAPTER 10

CURRENT STUDY

The current study combines computational modeling and neuroimaging experiment to investigate the neural mechanisms of pronoun resolution in English and Chinese. Given the typological difference in English and Chinese pronouns, we hypothesize that Chinese speakers rely more on discourse-level information during pronoun resolution, whereas English speakers are more sensitive to morpho-syntactic constraints.

The lack of overt morphological marking on Chinese pronouns and the *pro*-drop phenomenon leads to a typological difference in Chinese and English sentence structures, namely, Chinese is a topic-prominent language with a topic-comment structure, whereas English is a subject-prominent language that has a subject-predicate structure (Chapter 3). One consequence for this typological difference is that Chinese speakers attend more to the topic, or the most prominent entity in the current discourse context during sentence comprehension, and are hence more likely to link the most prominent entity with a following pronoun, consistent with salience-based accounts on pronoun resolution [see Section 2.2 7, 39]. On the other hand, the English speakers have to match gender and number information between the antecedent and the pronouns, thus they may be more sensitive to morpho-syntactic cues during pronoun resolution.

To examine this hypothesis regarding English and Chinese pronoun resolution, we correlated brain activity with complexity metrics derived from the three computational models introduced in Part III. The syntax-sensitive Hobbs algorithm contains a morphological agreement feature which is specific to English, and the ACT-R model is consistent with the salience-based account for pronoun

resolution, which is more accurate for pronoun resolution in Chinese (see Chapter 9). We predict different brain activation patterns associated with the two metrics in English and Chinese, where the Hobbs algorithm is associated with activation in the syntax- and morphology-related regions in the English group, and the ACT-R metric is correlated with activation in the discourse processing regions in the Chinese group.

We also examined brain activity associated with the neural coreference model [18]. This model differs from both the Hobbs and the ACT-R model and encompasses a large set of features including word embeddings and some discourse features such as linear distance between the pronoun and the antecedent. Trained on the CONLL-2012 Shared Task corpus [85], this model achieved state-of-the-art results with F1 scores of 65.39 and 63.66 for the English and Chinese test data in the corpus. The high performance of the neural network model makes it a possible cognitive model for pronoun resolution.

To further examine the status of the computational models as cognitive models for pronoun resolution, we also correlated brain activity with binary regressors that simply marks 1 at each first, second and third person pronoun. The activation map for the binary third person pronoun regressor indicates brain regions activated for the presence of third person pronouns, hence is expected to be the superset of the activation maps correlated with the three complexity metrics. In the next chapter we describe our neuro-computational modeling approach in detail and our fMRI data acquisition procedure.

CHAPTER 11

METHODS

The current study applies the neuro-computational modeling approach, pioneered by Brennan [11], to examine the neural mechanisms for pronoun resolution in English and Chinese. A neuro-computational model involves a computational psycholinguistic model and a linking hypothesis. The computational psycholinguistic model derives intermediate states of how specific syntactic, semantic, or other processes are engaged at specific words, and the linking hypothesis connects the states to observable neural signals. This chapter reviews the neuro-computational approach and the linking hypothesis applied to our pronoun resolution models, and described the fMRI data collection and analysis procedure.

11.1 Neuro-computational models

11.1.1 Overview of the approach

Neuro-computational models are built upon computational psycholinguistic models that quantifies how linguistic knowledge is deployed in real time. A computational psycholinguistic model operates over sequences of words and quantifies how a specific computation, such as a syntactic parse states, unfolds word-by-word during comprehension [see 41]. In the neuro-computational modeling approach, the incremental mental states derived by psycholinguistic computational models are further quantified by a linking hypothesis to estimate brain signals, which are tested against actual brain data. For instance, surprisal

of word-category probability based on a context-free grammar or the number of syntactic nodes could both serve as a complexity metric that links mental states to fMRI signals for syntactic processing [see 11].

Neuro-computational models can be tested against different types of brain data collected with different techniques in cognitive neuroscience, such as bold-oxygen-level-dependent (BOLD) data recorded by fMRI, and magnetic fields induced by current flow collected with magnetoencephalography (MEG). This model-based approach, coupled with naturalistic stimuli such as story listening, allows investigation of the location and timing of fine-grained linguistic processing in the brain without experimental manipulation.

Another advantage of the neuro-computational approach, compared with the traditional controlled experiments, is its ability to target distinct sub-processes of sentence comprehension in a natural setting. Wehbe et al. [111], for example, built a multi-faceted model of reading that consists of lexical semantic features, syntactic features like subject and object, and discourse-level features like character reference. They combined this complex model with fMRI data from story-reading and revealed an activation map where the traditional language comprehension network is divided into sub-regions specific to the lexical semantic, syntactic, and discourse-level features.

The current work applies the neuro-computational modeling approach to examine pronoun resolution. We first marked the offset of each word in the audiobook of "The Little Prince", and we computed the predicted mental states for pronoun resolution at each third person pronoun in the audiobook using the three computational models discussed in Part III. These states derived by the computational models are then transformed by their linking hypotheses to

quantify the processing difficulty of pronoun resolution, which are then convolved with the canonical hemodynamic response function to derive an estimate of the fMRI time-course during pronoun resolution. We then compared fits of the estimated BOLD signals with the BOLD signals that are recorded while participants passively listened to the audiobook in the fMRI scanner. If the fits between the estimated signals and the observed signal is high in a specific brain region, it indicates that this region may well reflects processing integrated in the computational psycholinguistic model for pronoun resolution. Figure 11.1 provides a concrete illustration of the neuro-computational approach for the current study.

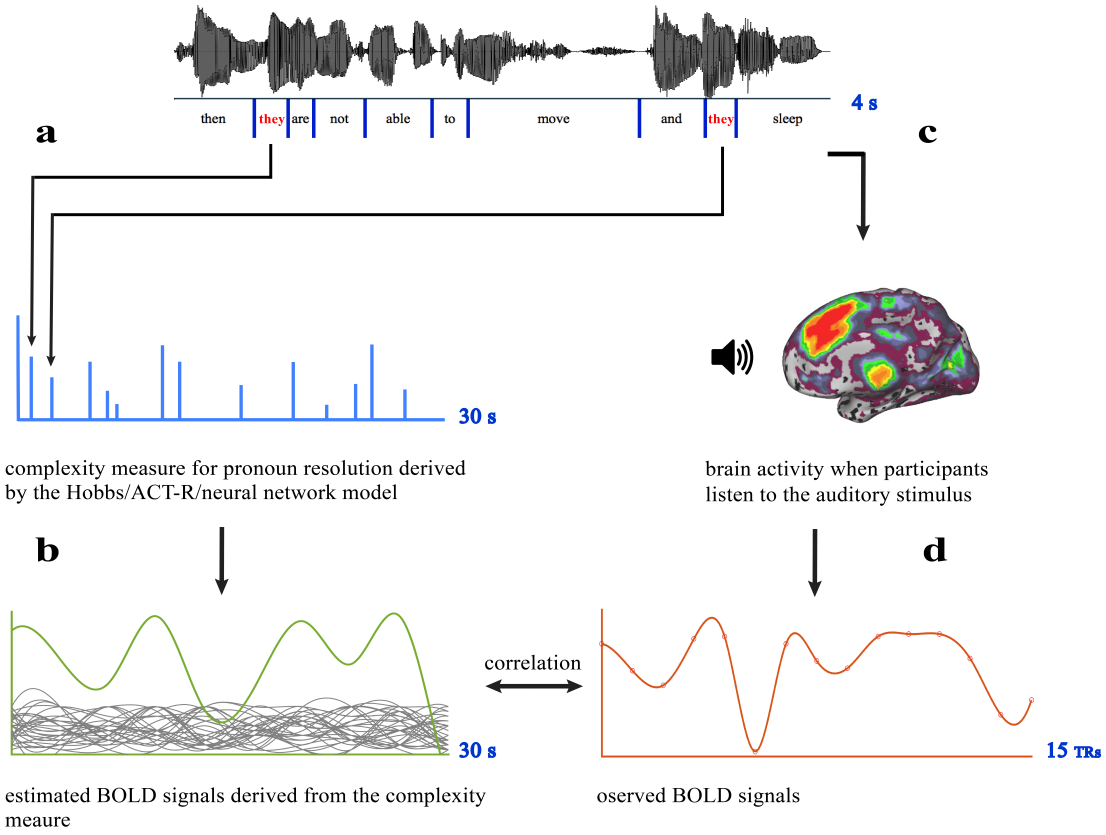


Figure 11.1: An illustration of the neuro-computational approach. The top of the figure shows a segment of the audiobook stimulus. Word boundaries are indicated in blue. (a) The complexity measures derived by the computational models quantify the processing difficulty of linking a pronoun to its antecedents at the offset of each pronoun in the auditory stimuli; (b) The complexity measures are convolved with the canonical hemodynamic response function to derive an estimate of the time-course of BOLD signals that reflect pronoun resolution difficulty. (c) The brain activity are recorded while participants passively listened to the audiobook. (d) The observed BOLD signals from a specific brain region is extracted and correlated against the estimated signal in (c) to test how well this brain region reflects processing difficulty in pronoun resolution derived by the complexity measures.

11.1.2 The linking hypothesis

As discussed in Part III, the output of the Hobbs, ACT-R and the neural network models for pronoun resolution are the Hobbs distance, the ACT-R activation level

and the neural coreference score, respectively. A higher Hobbs distance indicates more competing antecedents and more syntactic and morphosyntactic operations; a higher ACT-R activation level suggests higher salience of the antecedent in the discourse context based on recency, frequency and grammatical role of the antecedent, and a higher neural coreference score suggests higher probability for the antecedent and the pronoun to co-refer based on a set of semantic and discourse level information.

Given that increased hemodynamic response of a brain region indicates increased neural activity in that region, we need to first transform the outputs from the the computational models to complexity metrics in order to connect the models with brain activity. The Hobbs distance itself qualifies as a complexity metric as it is positively correlated with processing efforts. The ACT-R activation level and the neural coreference score are negatively correlated with difficulty in pronoun resolution so we took the negative of the two measures as their complexity metric.

The complexity metrics were used to derive estimated brain states, which are aligned with brain signals with a response function. A response function mediates between the actual physiological activity of neurons and the brain signals measured with a specific technique such as fMRI, MEG or EEG. For fMRI research, the complexity metrics are convolved with the canonical hemodynamic response function (HRF) to account for the delay between neuronal activity and measured changes in blood oxygenation level. The complexity metric, together with the response function, constitutes a linking hypothesis that connects the properties of a theoretical mental state with an observable brain signal [23]. Table 11.1 lists the parameters of the neuro-computational models in the current

study. The third and the fourth column is the linking hypothesis.

Model	Output	Complexity metric	Response function
`Hobbs`	Hobbs distance	Hobbs distance	HRF
`ACT-R`	activation level	negative activation level	HRF
`Neural Coreference`	coreference score	negative coreference score	HRF

Table 11.1: Parameters in neuro-computational models of the current study.

11.2 Experiment

11.2.1 Participants

English participants are 49 healthy, right-handed, young adults (30 female, mean age = 21.3, range = 18-37). They self-identified as native English speakers, and had no history of psychiatric, neurological or other medical illness that could compromise cognitive functions. All participants were paid, and gave written informed consent prior to participation, in accordance with the guidelines of the Human Research Participant Protection Program at Cornell University.

Chinese participants are 35 healthy, right-handed, young adults (15 female, mean age=19.3, range = 18-25). They self-identified as native Chinese speakers, and had no history of psychiatric, neurological or other medical illness that could compromise cognitive functions. All participants were paid, and gave written informed consent prior to participation, in accordance with the guidelines of the Ethics Committee at Jiangsu Normal University.

11.2.2 Stimuli

The English audio stimulus is an audiobook version of Antoine de Saint-Exupéry's *The Little Prince*, translated by David Wilkinson and read by Nadine Eckert-Boulet. This text contains 3127 non-pronominal mentions and 645 first person pronouns, 302 second person pronouns and 675 third person pronouns (see Table 9.2).

The Chinese audio stimulus is a Chinese translation of *The Little Prince* [1], read by a professional female Chinese broadcaster. Within this text, 2947 non-pronominal mentions and 639 first person pronouns, 298 second person pronouns and 529 third person pronouns are identified (see Table 9.2).

11.2.3 Procedure

After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner. The presentation script was written in PsychoPy [82]. Auditory stimuli were delivered through MRI-safe, high-fidelity headphones (English: Confon HP-VS01, MR Confon, Magdeburg, Germany; Chinese: Ear Bud Headset, Resonance Technology, Inc, California, USA) inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. An experimenter increased the sound volume stepwise until the participants could hear clearly.

The English and Chinese audiobooks lasted for 94 and 99 minutes, respectively. They were both divided into nine sections, each lasted for about ten

[1] http://www.xiaowangzi.org

minutes. Participants listened passively to the nine sections and completed four quiz questions after each section (36 questions in total). These questions were used to confirm their comprehension and were viewed by the participants via a mirror attached to the head coil and they answered through a button box. The entire session lasted for around 2.5 hours.

11.2.4 MRI Data Collection and Preprocessing

Both English and Chinese brain imaging data were acquired with a 3T MRI GE Discovery MR750 scanner with a 32-channel head coil. Anatomical scans were acquired using a T1-weighted volumetric Magnetization Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence. Blood-oxygen-level-dependent (BOLD) functional scans were acquired using a multi-echo planar imaging (ME-EPI) sequence with online reconstruction (TR=2000 ms; TE's=12.8, 27.5, 43 ms; FA=77°; matrix size=72 x 72; FOV=240.0 mm x 240.0 mm; 2 x image acceleration; 33 axial slices, voxel size=3.75 x 3.75 x 3.8 mm). Cushions and clamps were used to minimize head movement during scanning.

All fMRI data were preprocessed using AFNI version 16 [21]. The first 4 volumes in each run were excluded from analyses to allow for T1-equilibration effects. Multi-echo independent components analysis (ME-ICA; [60]) were used to denoise data for motion, physiology and scanner artifacts. Images were then spatially normalized to the standard space of the Montreal Neurological Institute (MNI) atlas, yielding a volumetric time series resampled at 2 mm cubic voxels.

11.2.5 Statistical Analysis

A GLM analysis was conducted to compare the mechanisms for third pronoun resolution in English and Chinese. The BOLD signals was modeled by the complexity metrics derived from the Hobbs, the ACT-R and the neural coreference model time-locked at the offset of each third person pronoun in the audiobook (see Section 11.1 for a detailed description of the linking hypotheses and the complexity metrics). Only third person pronouns are included because they provide gender information in English but not in Chinese, which points to potentially different brain activation maps.

The binary third person pronoun regressor was also included as a control variable. The other three control variables are the same with the first GLM model: `RMS intensity`, `word rate` and `frequency`.

The full GLM for the Hobbs, ACT-R, and neural coreference complexity metrics are as follows:

$$\text{BOLD} \sim \text{intensity} + \text{wordrate} + \text{freq} + \text{3rd\_pron} + \text{hobbs} + \text{actr} + \text{neuralnet}$$

At the group level, the activation maps for the complexity metrics derived from the three computational models were computed using one sample *t*-test. The voxelwise threshold was set at $p \leq 0.05$ FWE, with an adequate voxel size ($k \geq 50$). Contrasts of the activation maps between the English and Chinese groups were examined by a factorial design matrix, and statistical threshold was also set at $p \leq 0.05$ FWE. The GLM analysis was performed using SPM12 [83].

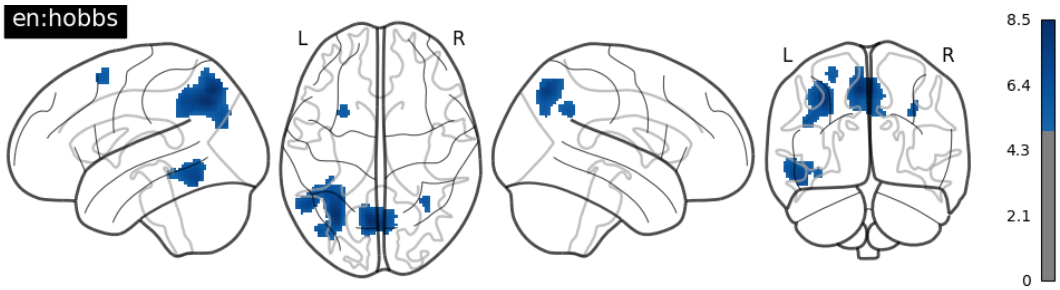
CHAPTER 12

RESULTS

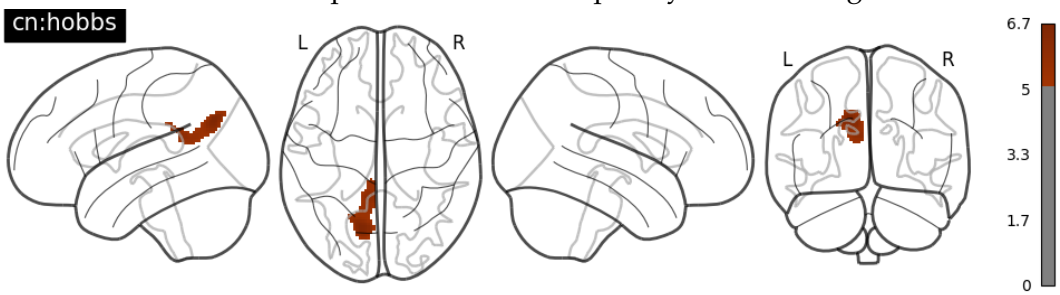
12.1 The Hobbs metric

Brain regions showing an increased activation for pronouns with higher processing difficulty predicted by the Hobbs complexity metric (i.e., the Hobbs distance) include the left Inferior Parietal Lobule (IPL), the left Precuneus cortex, the left ITG/MTG, the right AG and the left SFG for English speakers ($p < 0.001$ FWE, $k > 50$; see Figure 12.1a), whereas Chinese speakers have peak clusters in the left Precuneus cortex ($p < 0.05$ FWE, $k > 50$; see Figure 12.1b).

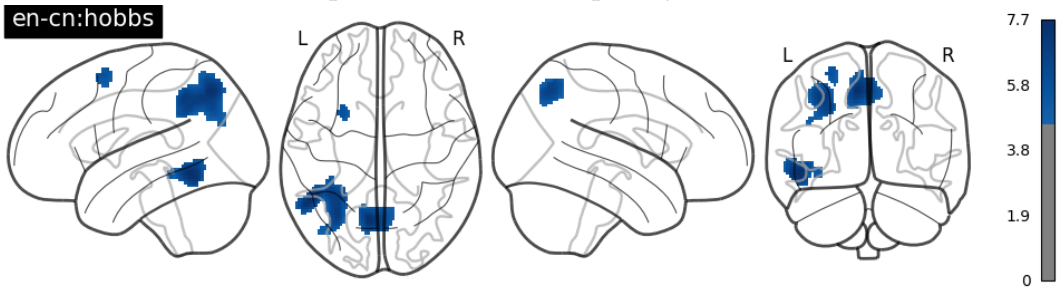
The difference between the Hobbs distance metric for third person pronoun resolution in Chinese and English is shown by the direct comparison reported in Table 12.1 ($p < 0.05$ FWE, $k > 50$). English speakers have stronger activation in the left ITG/MTG, the left Precuneus, the left IPL and the left MFG/SFG. Chinese speakers do not have stronger activity than English speakers for the Hobbs effects during third person pronoun resolution. Both English and Chinese speakers have significant activation for the Hobbs effect in the left Precuneus cortex (see Figure 12.1d).



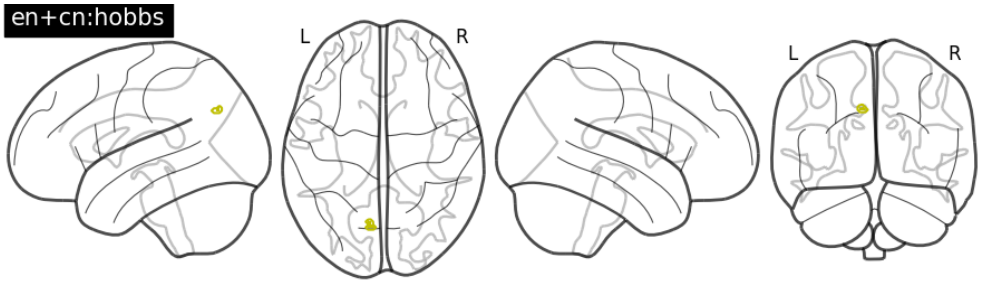
(a) T-score map for the Hobbs complexity metric in English



(b) T-score map for the Hobbs complexity metric in Chinese



(c) Contrast map of English greater than Chinese for the Hobbs complexity metric.



(d) Intersection map for the Hobbs complexity metric in English and Chinese.

Figure 12.1: Whole-brain effect with significant clusters for (a) the Hobbs complexity metric in English, (b) the Hobbs complexity metric in Chinese, (c) the English greater than Chinese contrast and (d) the intersection of English and Chinese Hobbs effect. The contrast map is inclusively masked for the positive effect of the Hobbs metric to avoid deactivation in the comparison. All images except for the intersection map underwent FWE voxel correction for multiple comparisons with $p < 0.05$.

Hobbs	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
complexity metric	x	y	z		*FWE-corr*	*cluster*	*peak*
English	-32	-64	42	left Inferior Parietal Lobule	< 0.001	1119	8.52
	-6	-66	50	left Precuneus	< 0.001	968	8.48
	-52	-56	-14	left Inferior/Middle Temporal Gyrus	< 0.001	406	7.18
	34	-54	34	right Angular Gyrus	0.003	72	5.91
	-26	12	60	left Superior Frontal Gyrus	0.004	72	5.86
Chinese	-14	-70	32	left Precuneus	0.001	432	6.67

(a) Significantly activated clusters by the Hobbs complexity metric for third person pronoun resolution in English and Chinese

Comparison of	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
Hobbs effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English > Chinese	-50	-52	-12	left Inferior/Middle Temporal Gyrus	< 0.001	390	7.69
	-8	-66	48	left Precuneus	< 0.001	788	7.61
	-32	-62	44	left Inferior Parietal Lobule	< 0.001	892	7.06
	-28	12	60	left Middle/Superior Frontal Gyrus	0.001	79	5.85

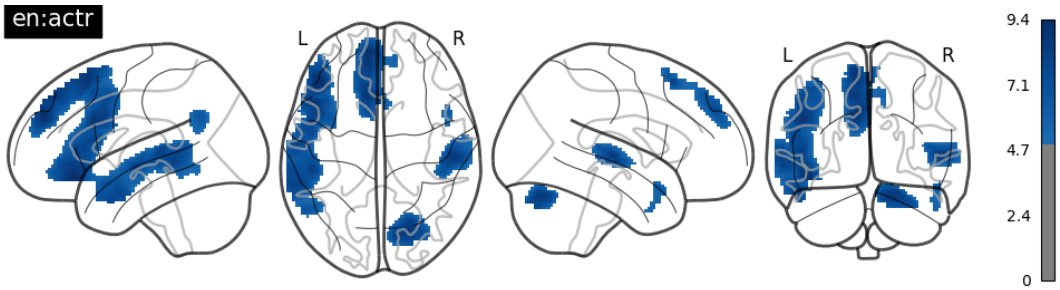
(b) Contrast between the Hobbs complexity metric for third person pronoun resolution in English versus Chinese.

Table 12.1: Significant clusters of BOLD activation for (a) the Hobbs complexity metric for third person pronoun resolution in English and Chinese and (b) their contrast after FWE voxel correction for multiple comparisons with $p < 0.05$ and $k > 50$. Peak activations are given in MNI Coordinates.

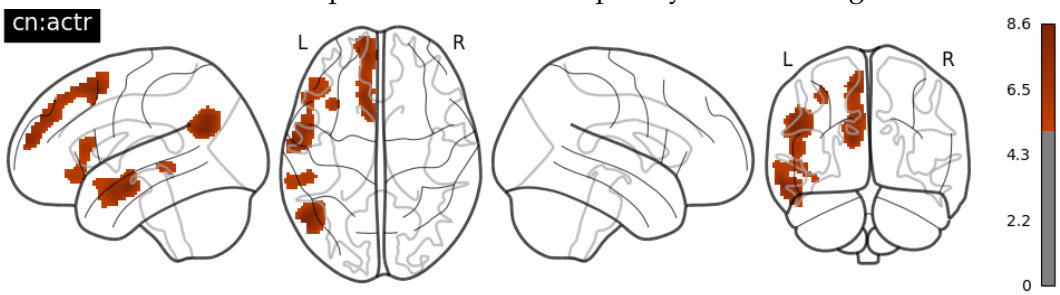
12.2 The ACT-R metric

For English speakers, the ACT-R complexity metric for third person pronoun resolution shows significant activation in the left IFG, the left SFG, the right Cerebellum, the right STG, the left AG and the right MTG ($p < 0.05$ FWE, $k > 50$; see Figure 12.2a). Chinese speakers have significant activation for the ACT-R complexity metric in the left AG, the left MTG, the left SFG and the left MFG ($p < 0.05$ FWE, > 50; see Figure 12.2b).

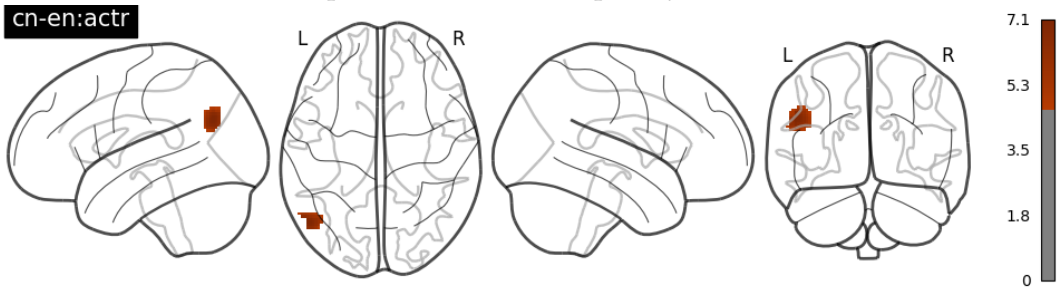
Direct comparison of the contrast maps for the ACT-R complexity metric reveals greater activity in the left AG for Chinese speakers. No significant activity is observed for English greater than Chinese at the corrected threshold ($p < 0.05$ FWE, $k > 50$; see Figure 12.2c). Table 12.2 lists the t-statistics for all the significant clusters. The region names are taken from the Harvard-Oxford Cortical Structure Atlas. The common brain regions that are associated with the ACT-R complexity metric for both English and Chinese include the left MTG, the left IFG, the left SFG and the left AG.



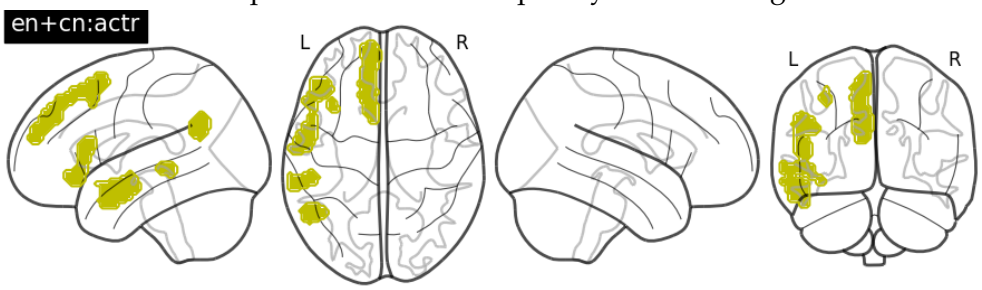
(a) T-score map for the ACT-R complexity metric in English



(b) T-score map for the ACT-R complexity metric in Chinese



(c) Contrast map for the ACT-R complexity metric in English and Chinese.



(d) Intersection map for the ACT-R complexity metric in English and Chinese.

Figure 12.2: Whole-brain effect with significant clusters for (a) the ACT-R complexity metric in English, (b) the ACT-R complexity metric in Chinese, (c) the Chinese greater than English contrast and (d) the intersection of English and Chinese ACT-R effect. The contrast map is inclusively masked for the positive effect of the ACT-R metric to avoid deactivation in the comparison. All images except for the intersection map underwent FWE voxel correction for multiple comparisons with $p < 0.05$.

ACT-R	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
complexity metric	x	y	z		*FWE-corr*	*cluster*	*peak*
English	-44	32	-14	left Inferior Frontal Gyrus (pars triangularis)	< 0.001	5826	9.41
	-8	54	30	left Superior Frontal Gyrus	< 0.001	1875	9.24
	20	-74	-26	right Cerebellum	< 0.001	740	8.67
	54	-22	2	right Superior Temporal Gyrus	< 0.001	780	8.53
	-54	-60	28	left Angular Gyrus	< 0.001	276	6.72
	52	12	-20	right Middle Temporal Gyrus	< 0.001	93	6.57
Chinese	-54	-64	24	left Angular Gyrus	< 0.001	644	8.61
	-60	-6	-18	left Middle Temporal Gyrus	< 0.001	650	8.36
	-10	64	16	left Superior Frontal Gyrus	< 0.001	903	8.19
	-34	20	50	left Middle Frontal Gyrus	0.001	68	6.68
	-52	24	10	left Inferior Frontal Gyrus (pars triangularis)	0.001	293	6.59
	-52	24	10	left Middle/Superior Temporal Gyrus	0.007	145	6.02

(a) Significantly activated clusters by the ACT-R complexity metric for third person pronoun resolution in English and Chinese

Comparison of	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
ACT-R effect	x	y	z		*FWE-corr*	*cluster*	*peak*
Chinese > English	-52	-62	28	left Angular Gyrus	< 0.001	200	7.07

(b) Contrast between the ACT-R complexity metric for third person pronoun resolution in English versus Chinese.

Table 12.2: Significant clusters of BOLD activation for (a) the ACT-R complexity metric for third person pronoun resolution in English and Chinese and (b) their contrast after FWE voxel correction for multiple comparisons with $p < 0.05$, $k > 50$. Peak activations are given in MNI Coordinates.

12.3 The neural coreference metric

The neural coreference complexity metric for third person pronoun resolution is associated with significant activation in the bilateral STGs for English speakers ($p < 0.001$ FWE, $k > 50$; see Figure 12.3). No significant clusters are found for the neural coreference metric in Chinese speakers at the corrected threshold.

Direct comparison between the two groups for the neural coreference metric reveals no region for either English greater than Chinese or Chinese greater than English at the corrected threshold ($p < 0.001$ FWE, $k > 50$). Table 12.3 lists the t-statistics for all the significant clusters using region names from the Harvard-Oxford Cortical Structure Atlas.

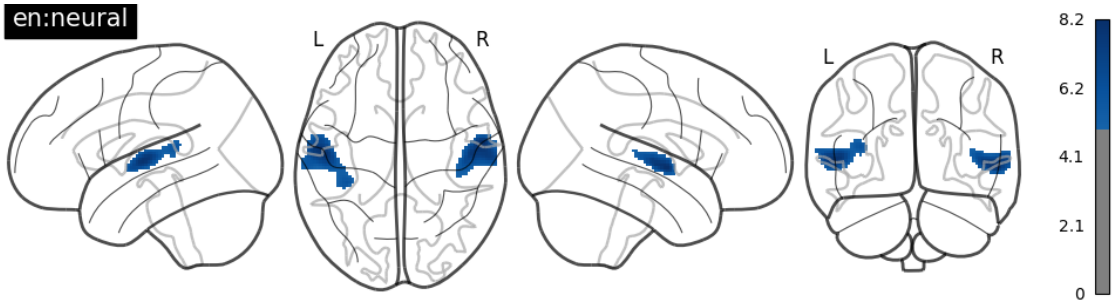


Figure 12.3: Whole-brain effect with significant clusters for the neural coreference complexity metric in English. All images underwent *FWE* voxel correction for multiple comparisons with $p < 0.05$.

Neural Network	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
complexity metric	x	y	z		*FWE-corr*	*cluster*	*peak*
English	56	-6	-2	right Superior Temporal Gyrus	< 0.001	570	8.21
	-54	-12	2	left Superior Temporal Gyrus	< 0.001	665	8

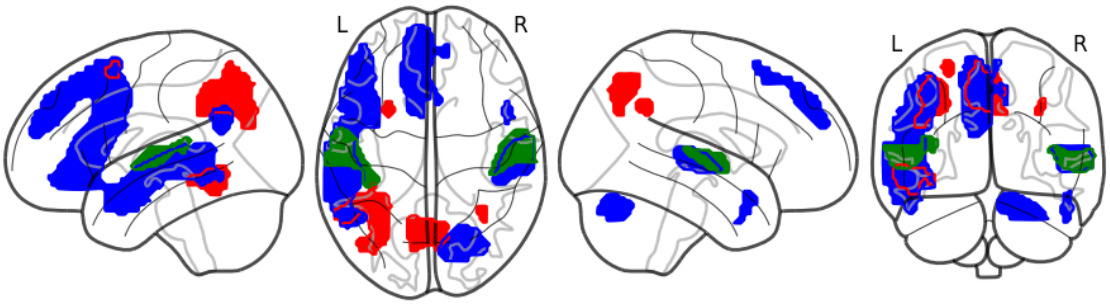
Table 12.3: Significant clusters of BOLD activation for the neural network complexity metric for third person pronoun resolution in English after *FWE* voxel correction for multiple comparisons with $p < 0.05$, $k > 50$. Peak activations are given in MNI Coordinates.

12.4 All complexity metrics

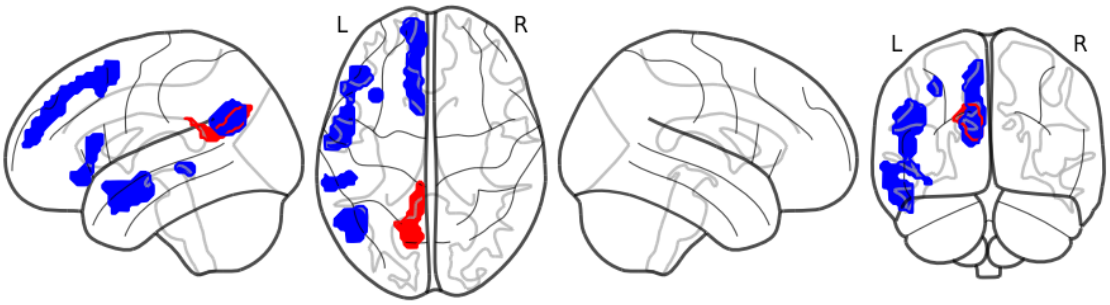
To sum up, the brain regions associated with the three complexity metrics reveal different activity patterns: the Hobbs distance predicts the left Precuneus activity in both English and Chinese, and is additionally associated with the left IPL, the left MTG and the left MFG activity in English. The negative ACT-R activation level predicts significant activity in the left AG, the left IFG and the left MFG for both English and Chinese, with stronger activity in the left AG in Chinese. The negative neural coreference score is associated with the bilateral STGs in English. Table 12.4 lists all the brain regions associated with the three complexity metrics. Figure 12.4 shows the activation map for the average main effects of the three complexity models.

Region	**Model**			
	English	Chinese	English > Chinese	Chinese > English
Left Inferior Parietal Lobule	Hobbs		Hobbs	
Left Angular Gyrus	ACT-R	ACT-R		ACT-R
Left Precuneus	Hobbs	Hobbs	Hobbs	
Left Middle Temporal Gyrus	Hobbs	ACT-R	Hobbs	
Left Middle Frontal Gyrus	Hobbs, ACT-R	ACT-R	Hobbs	
Left Inferior Frontal Gyrus (pars triangularis)	ACT-R	ACT-R		
Left Superior Temporal Gyrus	Neural	ACT-R		
Right Angular Gyrus	Hobbs			
Right Superior Temporal Gyrus	ACT-R, Neural			
Right Middle Temporal Gyrus	ACT-R			
Right Cerebellum	ACT-R			

Table 12.4: Summary of brain regions associated with for the Hobbs, ACT-R and neural coreference complexity metrics for third person pronoun resolution in English and Chinese.



(a) Overlays of the Hobbs, ACT-R and neural coreference complexity metrics for third person pronoun resolution in English.



(b) Overlays of the Hobbs and ACT-R complexity metrics for third person pronoun resolution in Chinese.

Figure 12.4: Overlays of the Hobbs, ACT-R and neural coreference complexity metrics for third person pronoun resolution in (a) English and (b) Chinese. Red color represents the Hobbs complexity effects; blue color represents the ACT-R complexity effects and green represents the neural coreference complexity effects. All images underwent FWE voxel correction for multiple comparisons with $p < 0.05, k > 50$.

CHAPTER 13

DISCUSSION

Activation maps of the three complexity metrics for third person pronoun resolution offer insights on the functions of different brain regions in the network of third person pronoun effects. The following subsections discussed six brain regions in detail: the left IPL, MTG, Precuneus, AG, IFG and STG.

13.1 Syntactic processing and the IPL

The Hobbs distance is associated with increased activity in the left IPL. Given that a higher Hobbs distance indicates more processing steps in the syntactic trees, one cognitive consequence is that the parser needs to hold more structural representations in the current mental state. We therefore relate the IPL activity with holding multiple structural representations, consistent with the neurolinguistic literature that associated IPL with syntactic ambiguity and non-canonical word orders.

The IPL activity has been reported in Tyler et al. [103] where participants listened to sentences containing a syntactically ambiguous phrase such as "bullying teenagers" in "The newspaper reported that bullying teenagers are a problem for the local school" or "The newspaper reported that bullying teenagers is bad for their self-esteem". Before the disambiguating predicate *are/is*, "bullying teenagers" is syntactically ambiguous as it can either be a complex noun phrase or a gerund verb construction. The fMRI results showed significant activity in the fronto-parietal region including the left IPL. Tyler et al. [103] therefore

suggested that the IPL activity may reflect increased processing requirements involved in maintaining multiple syntactic representations. In another fMRI study, Fiebach et al. [27] varied the length of the syntactically ambiguous region where the disambiguating phrase either occurs early or late in the sentence. The results showed that the IPL activation positively correlated with the length of the syntactically ambiguous region.

Apart from syntactic ambiguity, IPL has also been associated with non-canonical word orders. Bornkessel-Schlesewsky et al. [10] asked the participants to judge the acceptability of German sentences where either the subject precedes the object (canonical) or the object precedes the subject (non-canonical). The results showed a left fronto-parietal network including the left IPL for the word order effect. Bornkessel-Schlesewsky et al. [10] argued that the left IPL is sensitive to syntactic reanalysis, consistent with the hypothesis that holding multiple syntactic representations activated the left IPL. Additionally, Yokoyama et al. [114] found that compared to lists of verbs or nouns, both active and passive sentences in English and Japanese activated the left IPL, among other fronto-temporal language regions.

Taken together, these findings all suggest the left IPL's involvement in maintaining syntactic representations. It is therefore unsurprising that we found the IPL activity significantly correlated with the Hobbs distance in English. However, there is no significant IPL activity for the Hobbs distance in Chinese, suggesting that pronoun resolution in Chinese may be less sensitive to morpho-syntactic processing burden.

13.2 Morphological processing and the left MTG

Another brain region that is associated with the Hobbs distance and is more activated in English compared to Chinese is the left MTG. A wealth of fMRI studies have implicated the left MTG in morphological processing such as gender and number matching, therefore, the contrast between the English and Chinese results suggests that English speakers are more engaged in morphological processing during pronoun resolution. Indeed, the gender/number checker only exists in the Hobbs algorithm for English but not for Chinese pronoun resolution.

As shown in Hammer et al. [42], German sentences with congruent biological (e.g., female) and syntactic gender (e.g., feminine marking) as in "Die Frau$_{female}$ ist beliebt, weil sie$_{female/feminine}$ schön ist."("The woman is popular, because she/she is beautiful.") were correlated with increased activation in the left temporal regions including the MTG and the STG, supporting this region's role in integrating biological and syntactic gender information during pronoun processing. Similarly, Miceli et al. [74] reported increased activation of the left MTG and the IFG in an grammatical gender judgment task where the subjects were asked whether a written noun has a masculine or feminine grammatical gender. Based on findings from lesion studies where left MTG lesion typically leads to aphasic patients with selective difficulty in accessing nouns [e.g., 62, 73], Miceli et al. [74] suggested that the left MTG is relevant for grammatical gender processing because grammatical gender is a property of nouns.

The recruitment of MTG in syntactic gender processing has also been found during language production. In a picture-naming task, Heim et al. [47] asked the participants to produce the definite determiner of the objects in German,

which requires grammatical gender selection (masculine/feminine). Compared to a naming task where the participants do not produce the determiner, gender selection elicited greater activation in the left IFG, STG and MTG. Based on a comprehensive reviews of neuroimaging studies on syntactic gender processing, Heim [46] further proposes a neural model where gender information is stored in the left temporo-parieto-occipital junction or the left MTG. From there it is retrieved in the left IFG pars triangularis and evaluated in the left IFG pars opercularis. This model is consistent with Indefrey and Levelt's [56] model on language production where the left MTG is the site for lemma selection and retrieval, including accessing syntactic gender information.

Apart from gender processing, the left MTG has also been associated with morphological inflection in general. Compared to stems of nouns and verbs such as "snail" and "hear", plurals and inflected verbs such as 'snails" and "hears" elicited greater activation in the left MTG and left IFG [67]. Thus the left MTG may be involved in morphological processing in general, including both gender and number matching.

Given converging evidences on the role of the left MTG in morphological processing such as accessing gender and number information, our finding that the left MTG is more activated in English than in Chinese further confirms the importance of morpho-syntactic cues during pronoun resolution in English.

13.3 Reference tracking and the left Precuneus

The Hobbs distance is also correlated with increased activity in the left medial parietal lobe/Precuneus for both English and Chinese. This medial parietal acti-

vation has been repeatedly reported in the neuroimaging literature on pronoun resolution when there are multiple referents in the discourse context.

Nieuwland et al. [79], for example, compared BOLD responses to sentences containing referentially ambiguous pronouns, as in "Ronald told Frank that he ..." and sentences with referential coherent pronouns, as in "Ronald told Emily that he ...". The results suggested that referential ambiguity selectively recruited the left medial parietal region/Precuneus. In another fMRI study, Boiteau et al. [9] found increased parietal activity associated with sentences involving two referents (e.g., "Jeremy and Lucy did some work on the house next door.") compared to sentences containing only one singular subject (e.g., "Jeremy did some work on the house next door."). Similarly, McMillan et al. [72] found widespread frontal-parietal activity including the left Precuneus associated with ambiguous pronominal reference than non-ambiguous pronominal reference. Brodbeck and Pylkkänen [14], using a visual world paradigm in MEG, found that successful reference resolution associated with increased activity in the medial parietal lobe. Wehbe et al. [111], using an integrated computational modeling approach, also identified the Precuneus cortex for the sub-processes of tracking characters during narrative reading.

Taken together, these findings all suggest that the left Precuneus may be responsible for the maintenance and integration of multiple representations. The parietal involvement in maintaining and integrating references is also supported in Almor et al.'s [2] finding that repeated names (e.g., "Susan is really into animals. Susan gave Betsy a pet hamster.") engaged greater temporal and parietal activity than pronouns (e.g., "Susan is really into animals. The other day she gave Betsy a pet hamster."). Since repeated names are associated with

the temporary addition of a new discourse entity before it is resolved as being coreferential, increased parietal activity therefore reflects recruitment of circuits tracking multiple discourse referents. Almor et al. [2] further suggested that the parietal regions might be originally devoted to perceptual organization where it tracks multiple objects in space.

Furthermore, neuroimaging studies on decision-making has implicated the parietal cortex in the integration of the components contributing to probability and risk of choosing an outcome [e.g., 53, 71, 108]. This led McMillan et al. [72] to argue that pronoun resolution involves a decision-making mechanism where the comprehenders strategically choose a pronoun's referent in a probabilistic manner that maximizes the likelihood of correctly identify the referent and minimizes the risk of misinterpreting a sentence.

The parietal regions has also been associated with discourse coherence. According to Ferstl et al.'s [26] meta-analysis of neuroimaging studies on text comprehension, the processing of coherent language is associated with increased activity in medial parietal, medial frontal and bilateral temporal areas. Kuperberg et al. [61] also reported increased activation of the parietal cortex during the processing of semantically unrelated sentences. Therefore, the parietal regions are relevant for referential processing since referential coherence is a fundamental component of discourse coherence.

To sum up, there is converging evidence for the recruitment of the left medial parietal region/Precuneus during the processing of referentially ambiguous pronouns. This directly relates to our finding that the left Precuneus is only significant for the Hobbs distance: the Hobbs distance indicates the number of proposals that the algorithm skips until it reaches the correct antecedent, and a

higher Hobbs distance indicates a larger number of competing antecedents in the discourse context.

One key result of our comparison between the English and Chinese groups is the stronger activity associated with the Hobbs distance in the left Precuneus for the English group. Since a higher Hobbs distance indicates an increased number of competing referents, the greater activation in the left Precuneus in English suggests that English speakers encounter more difficulty in processing referentially ambiguous pronouns.

13.4 Syntax-semantic integration and the left AG

The ACT-R complexity metric is associated with the left AG for both the English and Chinese groups, with greater activity in the Chinese group. Since the ACT-R metric includes both discourse-level information (recency and frequency of the antecedent) and syntactic information (grammatical role of the antecedents; see Table 9.1), we expect the ACT-R metric to be correlated with brain regions responsible for integrating information from both discourse and syntactic aspects of sentence comprehension. The left AG has been suggests to serve exactly this function [8].

Structurally, the AG adjoins the visual, spatial, auditory, and somatosensory regions. This makes it the best candidate for a high-level, supramodal integration area in the human brain [33]. Lesions in this region led to a variety of deficit in sentence comprehension [see e.g., 22]. Specifically, patients with lesions in the left AG made more "role reversal errors" (i.e., who did what to whom) in a sentence-picture matching task where they heard sentences either in the active

voice (e.g. "The horse chases the boy") or the passive voice (e.g., "The boy is chased by the horse") and had to match the sentence to the appropriate picture out of an array of three pictures [103]. This suggests that the left AG might be involved in processing Argument structure, which well explains the grammatical role elements in the ACT-R complexity metric for pronoun resolution.

The left AG has also been implicated in combinatorial conceptual/semantic processing. Humphries et al. [54], for example, observed the left AG activity for semantically congruent sentences compared to semantically incongruent sentences, semantically congruent word lists, random word lists and pseudo-word lists. In addition, the left AG activation occurred at the end of the sentences, indicating its function of combining elemental lexical concepts into a coherent discourse. This is consistent with a number of neuroimaging studies that also supported AG activation for words of higher combinatorial strength compared to unrelated words [86] and connected discourse compared to unrelated sentences [29, 50, 113]. Binder et al. [8] conducted a comprehensive meta-analysis on semantic processing and suggested that the AG is at the top of a processing hierarchy underlying concept retrieval and conceptual integration, and is therefore essential for discourse comprehension. Since pronoun resolution may also involve an integration process where the pronoun and the antecedent are combined, it may well engage the left AG activation for high-level conceptual processing.

In addition, when compared with *Wh* filler-gap constructions (e.g., "Which song$_i$ did the band play$_{}_i$ at the concert that ended early?"), backward anaphora (e.g., "Because he$_i$ extinguished the flames, the fireman$_i$ saved the resident that arrived later.") elicited more activation in the left AG [70]. This suggests that

pronoun-referent linking may involve a different mechanism than the filler-gap dependency. Given that the left AG has been associated with semantic processing, Matchin et al. [70] suggested that anaphora resolution might involve both semantic and syntactic processing whereas *Wh* constructions may rely mainly on syntactic processing or working memory/ cognitive control at the left IFG.

Considering these various lines of evidence supporting the left AG for higher-level syntactic and semantic integration, the greater activity we observed in the left AG for the ACT-R metric in Chinese indicates that Chinese speakers are utilizing both semantic/discourse and syntactic information during third person pronoun resolution. This supports our hypothesis that Chinese speakers rely more on contextual information to resolve the referent due to the lack of morphological cues.

13.5 Working memory, prominence and the left IFG

The ACT-R metric is also associated with the activation in the left IFG pars triangularis in both the English and Chinese groups. Given the distance feature contained in the ACT-R metric, the recruitment of the left IFG is not surprising as this region, especially the pars triangularis, has long been observed in tasks tapping working memory. The observed IFG activity is also consistent with Anderson [3]'s proposal that the declarative working memory module in ACT-R is associated with the left prefrontal region.

A number of studies have reported a distance effect where increased linear distance (i.e, number of words) between the filler and its gap in a sentence

produces greater activation in the region of Broca's area [e.g., 20, 28, 70]. Matchin et al. [70], for example, manipulated the linear distance between a back anaphora and its referent, and the distance between the filler and the gap. The distance effect revealed activity in the pars triangularis of the left IFG for both the back anaphora and the *Wh* constructions. Santi and Grodzinsky [92] also found a main effect of binding between antecedents and reflexive pronouns (e.g., "The girl supposes the cunning man hurt himself.") in the left IFG pars triangularis. Taken together with Matchin et al.'s finding, it seems that the left IFG, especially the pars triangularis region, is sensitive to working memory demand irrespective of different syntactic dependencies.

In addition to working memory, the left IFG has also been argued to be sensitive to the aboutness-based sequencing in sentence-initial positions [**?**]. The notion of "aboutness" generally defines the topic of the sentence. For example, in "John accused Mary of murder", the event is construed as being more strongly *about* John, whereas in "Mary was accused of murder by John" the event is more related to Mary. **?**] argued that in German, arguments in the sentence-initial position, whether it is a subject (i.e., *John*) or an object (i.e., *Mary*), is interpreted as the topic of the sentence. They compared brain activity correlated with subject-initial or object-initial word order in sentence-initial and sentence-medial sentence types, and the results showed increased activation in the pars opercularis of the left IFG for word order only, and increased activation for both word order and sentence type in the pars triangularis of the left IFG. **?**] then suggested a functional dissociation in the left IFG, with the pars opercularis is more related to sentence-internal prominence considerations, such as thematic roles, animacy, etc., and the pars triangularis supports discourse-level aboutness such as topical information.

Overall, the distance and the grammatical role features in the ACT-R metric are supported in left Broca's area which has been associated with working memory demands and topical information processing. There is no difference in this respect in the English and Chinese groups.

13.6 Semantic processing and the STGs

Lastly, we consider the bilateral STG's activation in response to the neural network's metric for English speakers. The superior temporal regions have long been associated with spectro-temporal analysis and prosodic processing in the language network [e.g., 48], yet they are also implicated in several investigations on semantic processing during sentence comprehension. For example, the anterior STG has been suggested to be particularly involved in processing the meaning of auditory and verbal stimuli [95, 98, 109].

Alternatively, the STG's structural connection to the IFG also makes it an ideal site for the integration semantic and syntactic information [31]. In the context of the current study, the bilateral STG activation well reflects the neural coreference metric, as the word embedding features of the neural network are meant to capture some lexical semantic information of the pronoun and the antecedent.

The neural network metric has no significant clusters in Chinese, this might imply a different strategies for pronoun resolution in English and Chinese. However, the null results could simply reflect the poor performance of the neural coreference model in Chinese (see Table 9.4). A better neural network model for Chinese pronoun resolution is needed for a meaningful comparison of pronoun resolution based on the features in the neural network model.

13.7 Towards a functional neuroanatomy of pronoun resolution

The presence of third person pronoun resolution elicites a network of activation in the fronto-temporal regions. The complexity metrics for third person pronoun resolution dissociate different regions in the network for different functions during third person pronoun resolution.

Specifically, the left IPL monitors and maintains multiple syntactic representations; the left MTG subserves morphological processing such as gender and number matching; the left Precuneus keeps track of referents in the current discourse context; the left AG supports higher-level integration of syntactic and semantic/discourse information; the left IFG is sensitive to working memory demands and discourse prominence of the referents, and the left STG is underlying integration of semantic information.

These regions connect and interact with each other during pronoun resolution, and based on language-specific factors needed in pronoun resolution, some regions may be recruited more than others in one language. For example, the left IPL and MTG for syntactic and morphological processing are more activated in English than in Chinese, whereas the left AG for higher-level integration is more activated in Chinese than in English. These different activity patterns support different processing mechanisms for pronoun resolution in English and Chinese due to typological differences of their pronouns.

Part V

Conclusion

CHAPTER 14

CONCLUSION

14.1 Summary of results

The current study investigates the neural mechanisms for pronoun resolution in English and Chinese by correlating complexity metrics derived from three computational models with brain activity during naturalistic story listening. Since Chinese pronouns lack overt gender and number marking in their spoken forms, we hypothesize that during pronoun resolution Chinese speakers would rely more on discourse-level information such as salience of the antecedent, whereas English speakers would additionally engage in morpho-syntactic processing.

This prediction is already borne out in the comparison of model performance in the English and Chinese text of *The Little Prince*: the syntax-sensitive Hobbs algorithm which includes syntactic locality constraints and morphological gender/number information performs better for third person pronoun resolution in English, whereas the discourse-based ACT-R model which focuses on salience of the antecedents is more accurate for Chinese third person pronoun resolution.

The correlational results with brain activity revealed different activation maps for pronoun resolution in English and Chinese, further confirmed the modeling results on the text data. The syntax-sensitive Hobbs algorithm is correlated with more activation in regions responsible for syntactic and morphological processing in English (the left IPL, Precuneus, MTG and MFG), whereas the discourse based ACT-R model has greater activation in Chinese the left AG which supports integration of syntactic information such as grammatical role

and lexical semantic information.

Compared with effect linked to the presence of third person pronouns, our model-based complexity metrics are able to disentangle different brain regions in the network of pronoun resolution and provides a clearer picture for the functional neuroanatomy of pronoun resolution. Specifically, we propose that the left IPL in the network is responsible for maintaining multiple syntactic representations of the relevant sentences during pronoun resolution, consistent with previous literature that also highlights this region in syntactic ambiguity [27, 103] and non-canonical word order [10, 114]; the left MTG subserves morphological processing such as gender and number matching during antecedent retrieval [42, 47, 74]; the left Precuneus is activated when there are competing antecedents in the context [9, 14, 72, 79], and the left AG supports higher-level integration of relevant syntactic and semantic information to achieve a coherent discourse comprehension [8, 22, 29, 50, 54, 86, 103, 113]. In addition, the left IFG recruitment indicates working memory demands during pronoun-antecedent linking [20, 28, 70, 92], and the left STG is involved for lexical semantic processing [95, 98, 109].

14.2 Significance

One key contribution of the current study is to disentangle different subprocesses in the complex process of pronoun resolution using computational models that focus on different aspects of pronoun resolution. Compared with previous studies on anaphora resolution that used experimental stimuli to investigate one or two factors in pronoun resolution such as distance [e.g., 70, 92] and

gender matching [e.g., 42] during pronoun resolution, our neuro-computational modeling approach is a first step towards a comprehensive picture of how different sub-processes of pronoun resolution is supported at the brain level.

This study is also the first to show a difference in the brain activation patterns during pronoun resolution for the English and Chinese speakers. The different activation maps support the hypothesized consequence for pronoun resolution in the two typologically distinct languages.

Furthermore, our study leverages naturalistic stimuli, which is especially important for pronoun resolution as in a natural setting, the relationship between antecedent and pronoun varies according to a large amount of linguistic, contextual and discourse information, which is usually absent from experimental stimuli. Our neuro-computational modeling approach, coupled with naturalistic stimuli, quantifies processing difficulty during pronoun resolution at each pronoun position in the discourse without any experimental manipulation.

14.3 Limitations

One unavoidable problem for the neuro-computational modeling approach is the different accuracy of the computational models selected. As shown in Table 9.4, the three computational models used in the current study differ greatly in their accuracy for third person pronoun resolution. This may lead to less accurate predictions of brain activity that correlated with the models. For example, the neural coreference model performs significantly worse for both the English and Chinese text. This is likely to explain why no significant cluster is observed for the neural coreference metric in Chinese, and its associated activation in English

might also be less reliable than the Hobbs and ACT-R metrics which have a much higher accuracy.

Another problem for the current study is the possible correlation among the features extracted from the three computational models. For example, the distance feature in the ACT-R model may largely correlate with structural distance which influences the Hobbs distance metric. There are also distance features in the neural network model, resulting in partial overlap in their predicted brain activity.

Thirdly, there are fewer Chinese participants (total number = 35) than English participants (total number = 49). This may lead to better estimate of the English activation patterns during pronoun resolution at the corrected threshold.

14.4 Future work

Given limitations on the performance of the models, especially the neural coreference model, one future direction is to first improve model performance. This is crucial for a conclusive conclusion for the involvement of hierarchical structures during pronoun resolution. Since the existing deep learning models on coreference resolution all include nominal and pronominal resolution, one possible way to improve the neural network model's performance is by training the model on pronouns only in the CONLL-2012 or the MUC-6 and MUC-7 corpora. Since most of the corpora contain newspaper articles, model adaptation may also be needed when applied to the novel genre.

A more accurate neural network model allows further analysis of brain

activity correlated with different layers in the neural network architecture, thus provides insights on the information flow in different brain regions during pronoun resolution. Currently, the vectors of the three layers in the mention-pair encoder of [18]'s neural coreference model with pre-trained weights do not differ much when applied to third person pronoun resolution in *The Little Prince*. With a more accurate neural network model, a ridge regression model described in Huth et al. [55] could be adapted to fit the different layers of the network with whole-brain activity.

The current study only shows whole brain activation pattern corresponds to the three complexity metrics. Future work may include ROI analyses of the brain regions based on previous neuroimaging studies on pronoun resolution (see Chapter 4) to further examine how the different elements in the models contribute to signal changes in the specific regions.

For the comparison between the process underlying English and Chinese pronoun resolution, there are other factors that are important for pronoun resolution and are not examined in the current study, such as animacy of the antecedent. As shown in Table 9.2, there are much more third person neutral pronouns *it* in English than in Chinese, suggesting that Chinese tends to avoid using pronouns for inanimate entities. This animacy feature could be integrated in the associative activation part of the ACT-R formula for a better prediction for Chinese pronoun resolution.

APPENDIX A

ADDITIONAL ANALYSES FOR PRONOUN EFFECTS

A.1 Methods

An additional GLM analysis was conducted to compare the effects of presence of the first, second and third person pronouns. At the single subject level, the observed BOLD time course in each voxel was modeled by the three binary regressors that simply mark 1 at the offset of each first, second and third person pronoun in the audiobook. We also added a binary regressor for non-pronominal mentions which marks 1 at the offset of each non-pronominal mention (see Section 9.2.1 for the count of non-pronominal mentions and the first, second, and third person pronouns in the English and Chinese text).

In addition, we included three control variables of non-theoretical interest: `RMS intensity` at every 10 ms of the audio; `word rate` at the offset of each spoken word in time; `frequency` of the individual words in Google Book unigrams [1]. These regressors were added to ensure that any conclusions about pronoun resolution would be specific to those processes, as opposed to more general aspects of speech perception.

The full GLM for the pronoun effects is as follows:

$$\text{BOLD} \sim \text{intensity} + \text{wordrate} + \text{freq} + \text{non\_pron} + \text{1st\_pron} + \text{2nd\_pron} + \text{3rd\_pron}$$

[1] http://books.google.com/ngrams

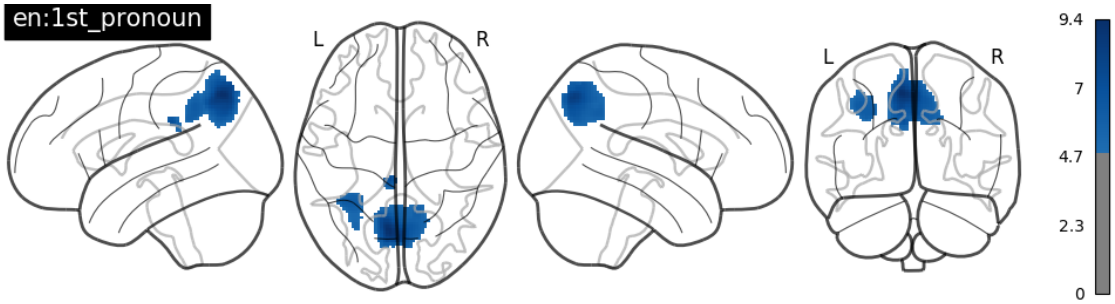
At the group level, the activation maps for the binary first, second and third person pronoun regressors in English and Chinese were computed using one sample t-test. The voxelwise threshold was set at $p \leq 0.05\ FWE$, with an adequate voxel size ($k \geq 50$). Contrasts of the activation maps between the English and Chinese groups were examined by a factorial design matrix, and statistical threshold was also set at $p \leq 0.05\ FWE$. The GLM analysis was performed using SPM12 [83].

A.2 Results

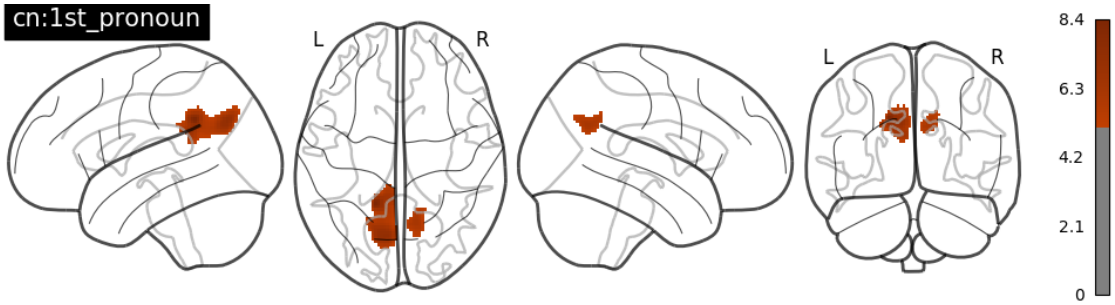
A.2.1 First person pronouns

The presence of first person pronouns is significantly associated with the left Precuneus, the left Cingulate Gyrus, and the left Supramarginal Gyrus/Angular Gyrus (AG) for English speakers. Similarly, for Chinese speakers, first person pronouns are associated with the left Cingulate Gyrus and the bilateral Precuneus cortex ($p < 0.001\ FWE$, $k > 50$; see Figures A.1a and A.1b).

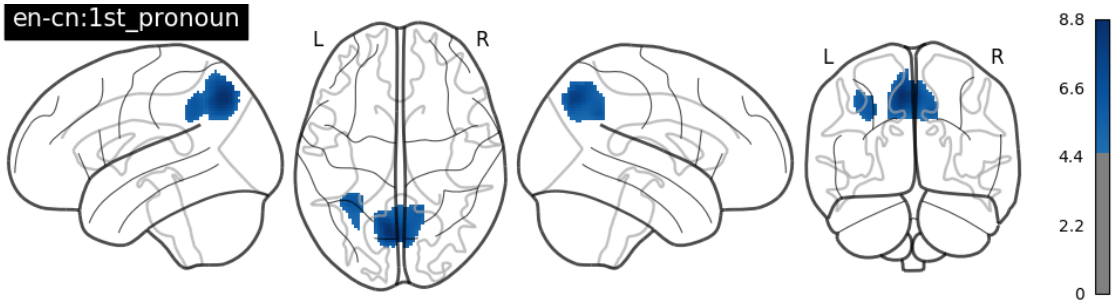
Direct comparison of the contrast maps between the English and Chinese groups suggests stronger activity in the left Precuneus and the left AG for English speakers ($p < 0.05\ FWE$, $k > 50$; see Figure A.1c). The contrasts are inclusively masked for the positive effect of first person pronouns to avoid deactivation in the comparison. No significant activity was found for Chinese greater than English for the first person pronoun effect. Table A.1 lists all the significant clusters and the t-statistics using region names from the Harvard-Oxford Cortical Structure Atlas.



(a) T-score map for the binary first person pronoun effect in English



(b) T-score map for the binary first person pronoun effect in Chinese



(c) Contrast map of English greater than Chinese for first person pronouns.

Figure A.1: Whole-brain effect with significant clusters for (a) binary first person pronouns effect in English, (b) binary first person pronouns effect in Chinese and (c) contrast map of English greater than Chinese for first person pronouns. The contrasts are inclusively masked for the positive effect of first person pronouns to avoid deactivation in the comparison. All images underwent FWE voxel correction for multiple comparisons with $p < 0.05$, $k > 50$.

1st person	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
pronoun effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English	-6	-68	48	left Precuneus	< 0.001	1998	9.35
	-6	-32	30	left Cingulate Gyrus	< 0.001	55	7.18
	-36	-48	40	left Supramarginal Gyrus/Angular Gyrus	0.001	377	6.34
Chinese	-16	-48	32	left Cingulate Gyrus/Precuneus	< 0.001	920	8.39
	12	-60	28	right Precuneus	0.004	164	6.21

(a) Significantly activated clusters by the binary first person pronoun effect in English and Chinese ($p < 0.05$ FWE, $k > 50$)

Comparison of	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
1st pronoun effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English > Chinese	-6	-68	46	left Precuneus	< 0.001	1649	8.78
	-34	-48	42	left Supramarginal Gyrus/Angular Gyrus	< 0.001	314	6.26

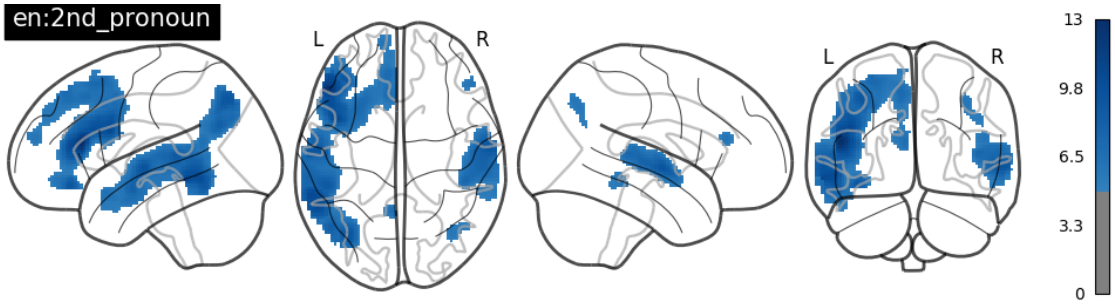
(b) Contrast between the first person pronoun effect in English versus Chinese ($p < 0.05$ FWE, $k > 50$)

Table A.1: Significant clusters of BOLD activation for (a) first person pronouns effects in English and Chinese and (b) the contrast of English greater than Chinese for first person pronoun effect after FWE voxel correction for multiple comparisons with $p < 0.05, k > 50$. The contrasts are inclusively masked for the positive effect of first person pronouns to avoid deactivation in the comparison. Peak activations are given in MNI Coordinates.

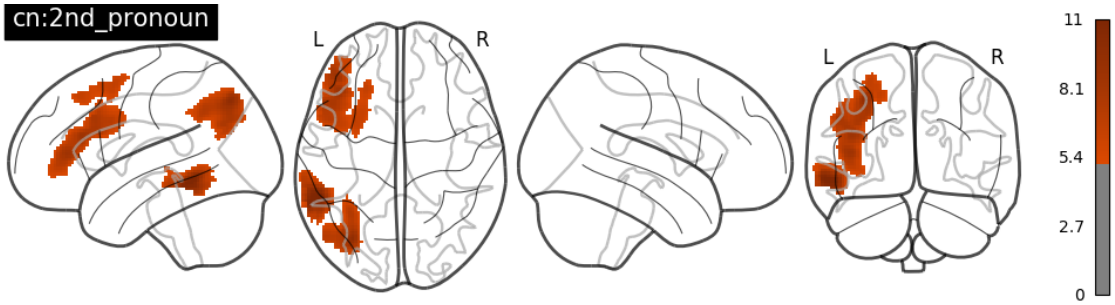
A.2.2 Second person pronouns

For English speakers, the presence of second person pronouns is significant in the bilateral Inferior Frontal Gyrus (IFG), the bilateral Middle Temporal Gyrus (MTG), the right Superior Temproal Gyrus (STG) and the left Precunes. The left MTG and IFG are also associated with second person pronouns in Chinese. The left Superior Frontal Gyrus (SFG) is significant for the presence of second person pronoun effect in Chinese ($p < 0.05$ FWE, $k > 50$; see Figures A.2a and A.2b).

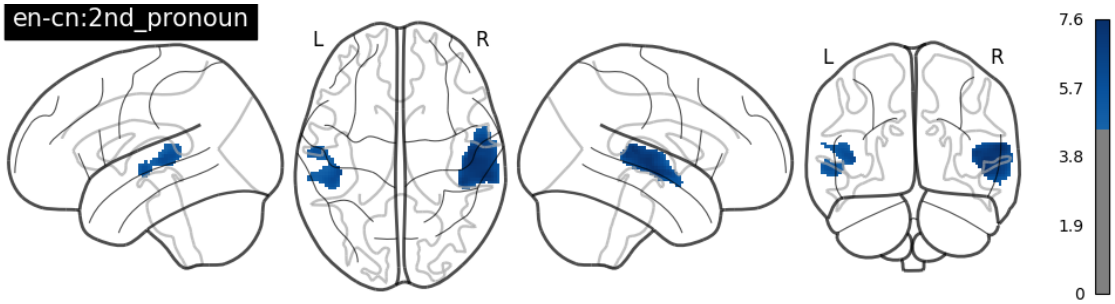
The contrast map shows activity in the right STG and the left MTG for English greater than Chinese. No stronger activity was found for Chinese greater than English for the presence of second person pronoun effect ($p < 0.05$ FWE, $k > 50$; see Figure A.2c). All the significant clusters and the t-statistics are given in Table A.2.



(a) T-score map for the binary second person pronoun effect in English



(b) T-score map for the binary second person pronoun effect in Chinese



(c) Contrast map of English greater than Chinese for second person pronouns.

Figure A.2: Whole-brain effect with significant clusters for (a) binary second person pronouns effect in English, (b) binary second person pronouns effect in Chinese and (c) contrast map of English greater than Chinese for second person pronouns. The contrasts are inclusively masked for the positive effect of first person pronouns to avoid deactivation in the comparison. All images underwent FWE voxel correction for multiple comparisons with $p < 0.05$, $k > 50$.

2nd person	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
pronoun effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English	-50	30	16	left Inferior Frontal Gyrus (pars triangularis)	< 0.001	4621	13.03
	-60	-42	-10	left Middle Temporal Gyrus	< 0.001	3633	10.48
	62	-10	-2	right Superior Temporal Gyrus	< 0.001	1421	9.4
	-6	-54	16	left Precuneus	< 0.001	83	6.72
	50	34	18	right Inferior Frontal Gyrus	0.001	58	6.22
	56	-38	-14	right Inferior/Middle Temporal Gyrus	0.004	89	5.82
Chinese	-54	-44	-12	left Inferior/Middle Temporal Gyrus	< 0.001	708	10.77
	-42	30	20	left Inferior Frontal Gyrus (pars triangularis)	< 0.001	1570	8.54
	-30	8	56	left Middle/Superior Frontal Gyrus	0.002	352	6.58

(a) Significantly activated clusters by the binary second person pronoun effect in English and Chinese ($p < 0.05$ FWE, $k > 50$)

Comparison of	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
2nd pronoun effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English > Chinese	50	-14	-4	right Superior Temporal Gyrus	< 0.001	1355	7.65
	-40	-26	4	left Middle Temporal Gyrus	< 0.001	321	6.21

(b) English greater than Chinese contrast for the second person pronoun effect ($p < 0.05$ FWE, $k > 50$)

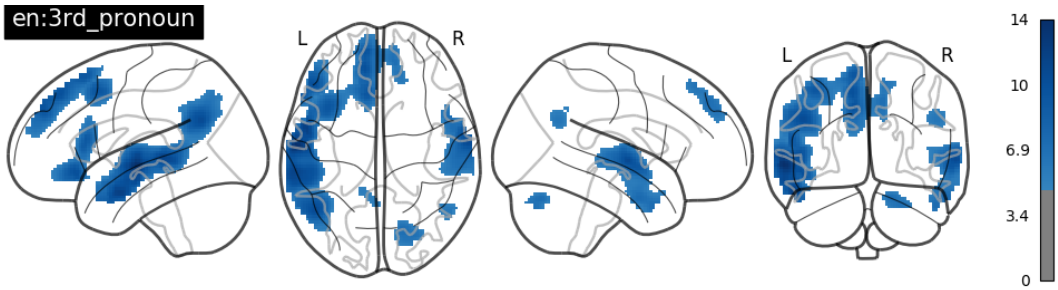
Table A.2: Significant clusters of BOLD activation for (a) second person pronouns effects in English and Chinese and (b) the contrast of English greater than Chinese for second person pronoun effect after FWE voxel correction for multiple comparisons with $p < 0.05$, $k > 50$. The contrasts are inclusively masked for the positive effect of second person pronouns to avoid deactivation in the comparison. Peak activations are given in MNI Coordinates.

A.2.3 Third person pronouns

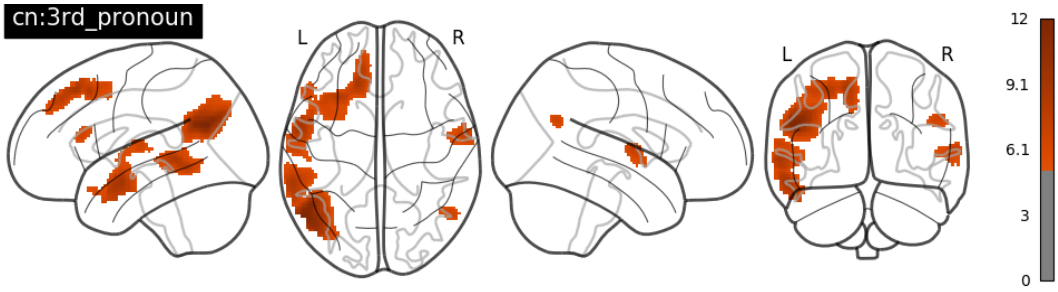
For English speakers, significant clusters associated with the presence of third person pronouns are observed in the bilateral STG, the left MTG, the left SFG, the left IFG, the right Cerebellum, the right AG and the left Precuneus ($p < 0.001$ FWE, $k > 50$; see Figure A.3a).

For Chinese speakers, the presence of third person pronouns is associated with increased activity in the bilateral AGs, the bilateral STGs, the left SFG, the left MTG and the left IFG ($p < 0.001$ FWE, $k > 50$; see Figure A.3b).

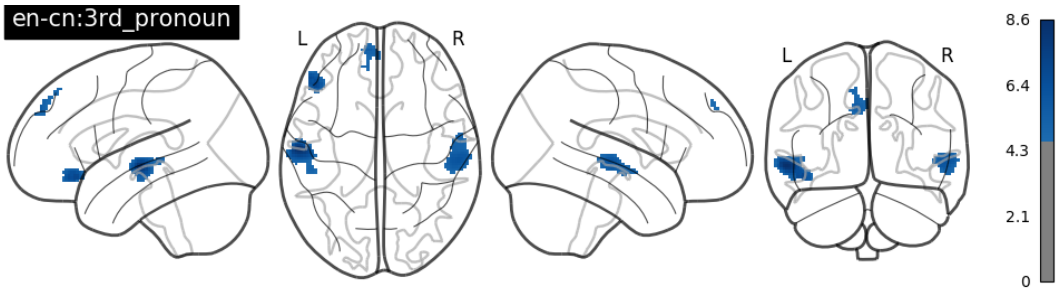
Direct comparison of the contrast maps between the English and Chinese groups suggests stronger activity in the bilateral STGs, the left IFG and the left SFG for English speakers. Chinese speakers showed stronger activity in the left AG than English speakers for the third person pronoun effect ($p < 0.05$ FWE; see Figures A.3c and A.3d). Table A.3 lists all the significant clusters and the t-statistics.



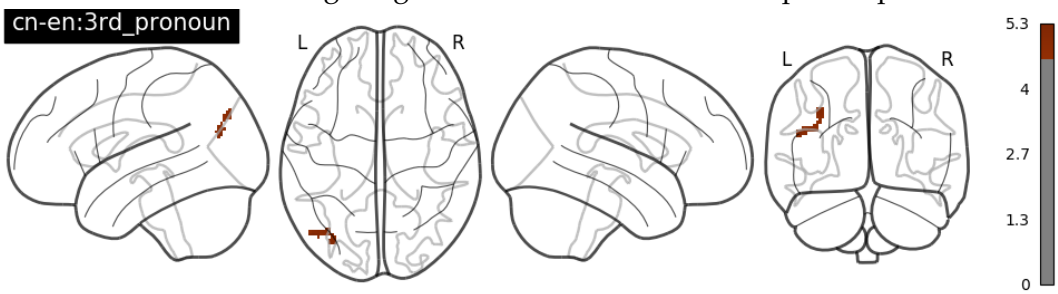
(a) T-score map for the binary third person pronoun effect in English



(b) T-score map for the binary third person pronoun effect in Chinese



(c) The contrast of English greater than Chinese for third person pronouns.



(d) The contrast of Chinese greater than English for third person pronouns.

Figure A.3: Whole-brain effect with significant clusters for (a) binary third person pronouns effect in English, (b) binary third person pronouns effect in Chinese, (c) the contrast map of English greater than Chinese and (d) the contrast map of Chinese greater than English. The contrasts are inclusively masked for the positive effect of second person pronouns to avoid deactivation in the comparison. All images underwent FWE voxel correction for multiple comparisons with $p < 0.05, k > 50$.

3rd person	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
pronoun effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English	-60	-12	-6	left Superior/Middle Temporal Gyrus	< 0.001	4476	13.75
	64	-10	-2	right Superior Temporal Gyrus	< 0.001	1660	10.96
	-10	42	46	left Superior Frontal Gyrus	< 0.001	2340	10.41
	-48	32	-10	left Inferior Frontal Gyrus (pars triangularis)	< 0.001	698	10.37
	18	-74	-30	right Cerebellum	< 0.001	241	7.2
	52	-60	28	right Angular Gyrus	0.001	127	6.3
	-12	-46	36	left Precuneus	0.008	73	5.6
Chinese	-52	-62	24	left Angular Gyrus	< 0.001	1531	12.12
	-58	-2	-18	left Middle/Superior Temporal Gyrus	< 0.001	791	9.19
	-14	32	50	left Superior Frontal Gyrus	< 0.001	848	8.88
	-62	-42	-2	left Middle Temporal Gyrus	< 0.001	665	8.66
	64	-6	6	right Superior Temporal Gyrus	< 0.001	211	7.29
	52	-62	28	right Angular Gyrus	0.001	94	6.74
	-52	26	16	left Inferior Frontal Gyrus (pars triangularis)	0.002	98	6.53

(a) Significantly activated clusters by the binary third person pronoun effect in English and Chinese ($p < 0.05$ FWE, $k > 50$)

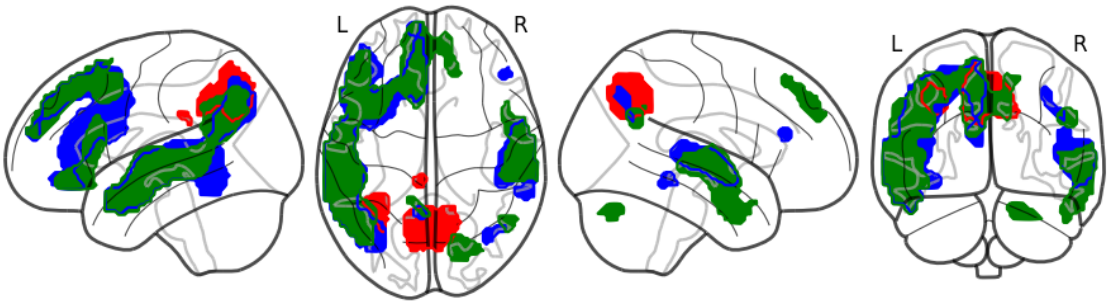
Comparison of	MNI coordinates			Region	*p*-value	*k*-size	*t*-score
3rd pronoun effect	x	y	z		*FWE-corr*	*cluster*	*peak*
English > Chinese	-58	-16	-6	left Superior/Middle Temporal Gyrus	< 0.001	391	8.6
	-46	30	-14	left Inferior Frontal Gyrus	< 0.001	102	7.08
	58	-22	-2	right Superior Temporal Gyrus	< 0.001	316	6.26
	-8	56	36	left Superior Frontal Gyrus	0.002	84	5.71
Chinese > English	-48	-74	20	left Angular Gyrus	0.006	54	5.33

(b) Contrast between the 3rd person pronoun effect in English versus Chinese ($p < 0.05$ FWE, $k > 50$)

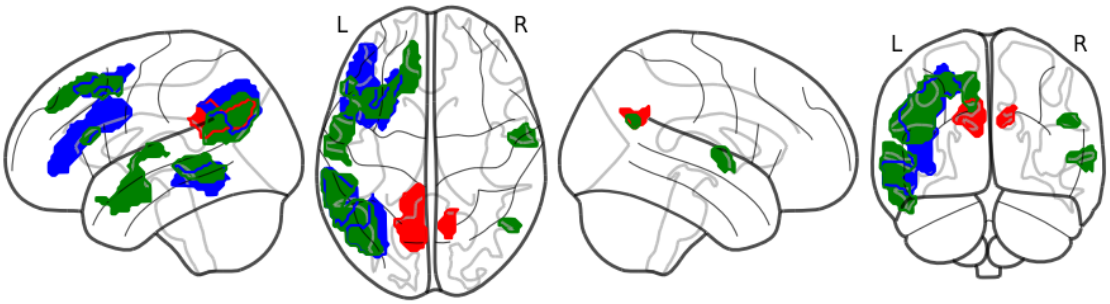
Table A.3: Significant clusters of BOLD activation for (a) third person pronouns effects in English and Chinese and (b) their contrast after FWE voxel correction for multiple comparisons with $p < 0.05$, $k > 50$. The contrasts are inclusively masked for the positive effect of second person pronouns to avoid deactivation in the comparison. Peak activations are given in MNI Coordinates.

A.2.4 All pronouns

First person pronouns highlight the Precuneus cortex while the second and third person pronouns are mainly associated with a left fronto-temporal network. There is no significant intersections among the three activation maps. Overlays of the average main effects for the first, second and third person pronoun effects in English and Chinese are shown in Figures A.4a and A.4b. Table A.4 lists all the brain regions associated with the presence of first, second and third person pronouns in English and Chinese.



(a) Overlays of first, second and third person pronoun effects in English.



(b) Overlays of first, second and third person pronoun effects in Chinese.

Figure A.4: Activation map for the average mean effects of the first, second and third person pronouns in (a) English and (b) Chinese. Red color represents first person pronoun effects; blue color represents second person pronoun effects and green represents third person pronoun effects. All images underwent FWE voxel correction for multiple comparisons with $p < 0.05, k > 50$.

Region	Pronouns			
	English	Chinese	English > Chinese	Chinese > English
Left Precuneus	1st, 2nd, 3rd	1st	1st	
Left Angular Gyrus	1st	3rd	1st	3rd
Left Cingulate Gyrus	1st			
Left Inferior Frontal Gyrus	2nd, 3rd	2nd, 3rd	3rd	
Left Middle Temporal Gyrus	2nd, 3rd	2nd, 3rd	2nd, 3rd	
Left Middle Frontal Gyrus			2nd	
Left Superior Temporal Gyrus	3rd			
Left Superior Frontal Gyrus	3rd	3rd	3rd	
Right Precuneus		1st		
Right Superior Temporal Gyrus	2nd, 3rd	3rd	2nd, 3rd	
Right Inferior Frontal Gyrus	2nd			
Right Middle Temporal Gyrus	2nd			
Right Angular Gyrus	3rd			
Right Cerebellum	3rd			

Table A.4: Summary of brain regions associated with first, second and third person pronouns in English and Chinese.

A.3 Discussion of the pronoun effects

One major difference for the first, second and third person pronoun effects is that the first person pronouns are associated with increased activation in the Precuneus cortex whereas the second and third person pronouns are left lateralized in a fronto-temporal network. The Precuneus activity has been assumed to support self-projection: the ability to mentally project oneself from the present moment into a simulation of another time, place, or perspective [**?**]. Additionally, the Precuneus has also been argued to be part of the default network that supports possession of a theory of mind, that is, understanding others' behavior from their perspectives [see **?** , for a meta-analysis]. In the current context, the Precuneus activity in the presence of first person pronoun during narrative understanding may well reflect a similar mentalizing mechanism where the participants project themselves to simulate the discourse characters' experience in order to understand them.

The presence of the second and third person pronouns elicited similar activity patterns in the left fronto-temporal network，including the left IFG, MTG and STG for both English and Chinese. Chinese has additional activation in the left AG for the third person pronoun effect, and English has greater activity in the right IFG and MTG for the second person pronoun effect. All these regions have been reported in previous neuroimaging studies that involves anaphora resolution (see Section 4.2). For the third person pronoun effect, the recruitment of the left IFG, MTG and STG supports pronoun resolution as a complex process that involves syntactic, morphological, semantic and discourse-level processing. The functional division of this network is evident in the results from our complexity metrics that target different aspects of pronoun resolution.

In the next section we discuss the relevant regions and their possible functions in the network of pronoun resolution.

BIBLIOGRAPHY

[1] Amit Almor and Veena A. Nair. The form of referential expressions in discourse. *Language and Linguistics Compass*, 1:84–99, 2007.

[2] Amit Almor, David V. Smith, Leonardo Bonilha, Julius Fridriksson, and Chris Rorden. What is in a name? Spatial brain circuits are used to track discourse references. *Neuroreport*, 18:1215–1219, 2007.

[3] John R. Anderson. Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, 29:313–341, 2005.

[4] John R. Anderson. *How can the human mind occur in the physical universe?* Oxford University Press, Oxford, 2007.

[5] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111:1036–1060, 2004.

[6] Mira Ariel. *Accessing noun-phrase antecedents*. Routledge, London, UK, 1990.

[7] Jennifer E. Arnold. How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4:187–203, 2010.

[8] Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19:2767–2796, 2009.

[9] Timothy W. Boiteau, Eric Bowers, Veena A. Nair, and Amit Almor. The neural representation of plural discourse entities. *Brain and Language*, 137: 130–141, 2014.

[10] Ina Bornkessel-Schlesewsky, Matthias Schlesewsky, and D. Yves von Cramon. Word order and Broca's region: Evidence for a supra-syntactic perspective. *Brain and Language*, 111:125–139, 2009.

[11] Jonathan Brennan. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10:299–313, 2016.

[12] Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J. Heeger, and Liina Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120:163–173, 2012.

[13] Susan Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, pages 155–162, Stroudsburg, PA, USA, 1987. Association for Computational Linguistics.

[14] Christian Brodbeck and Liina Pylkkänen. Language in context: Characterizing the comprehension of referential expressions with meg. *NeuroImage*, 147:447–460, 2017.

[15] Nancy A. Chinchor. Overview of MUC-7/MET-2. Technical report, Science Applications International Corporation, San Diego, 1998.

[16] Noam Chomsky. *Lectures on government and binding*. Foris, Dordrecht, Holland, 1981.

[17] Noam Chomsky. *Some concepts and consequences of the theory of government and binding*. MIT Press, Cambridge, Massachusetts, 1982.

[18] Kevin. Clark and Christopher D. Manning. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings*

of the 54th Annual Meeting of the Association for Computational Linguistics, volume 1, pages 643–653, Berlin, Germany, 2016. Association for Computational Linguistics.

[19] Kevin. Clark and Christopher D. Manning. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262. Association for Computational Linguistics, 2016.

[20] Ayanna Cooke, Zurif Edgar B., Christian DeVita, David Alsop, Phyllis Koenig, John Detre, James Gee, Maria Pinängo, Jennifer Balogh, and Murray Grossman. Neural basis for sentence comprehension: Grammatical and short-term memory components. *Human Brain Mapping*, 15:80–94, 2002.

[21] Robert W. Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, 29:162–173, 1996.

[22] Nina F. Dronkers, David P. Wilkins, Robert D. Van Valin, Brenda B. Redfern, and Jeri J. Jaeger. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92:145–177, 2004.

[23] David Embick and David Poeppel. Towards a computational(ist) neurobiology of language: Correlational, integrated, and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, 30:357–366, 2015.

[24] Gareth Evans. Pronouns. *Linguistic Inquiry*, 11:337–362, 1980.

[25] Murielle Fabre. *The sentence as cognitive object–The neural underpinnings of syntactic complexity in Chinese and French*. PhD thesis, INALCO Paris, 2017.

[26] Evelyn C. Ferstl, Jane Neumann, Carsten Bogler, and D. Yves von Cramon. The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, 29:591–593, 2008.

[27] Christian J. Fiebach, Sandra H. Vos, and Angela D. Friederici. Neural correlates of syntactic ambiguity in sentence comprehension for low and high span readers. *Journal of Cognitive Neuroscience*, 16:1562–1575, 2004.

[28] Christian J. Fiebach, Schlesewsky Matthias, D. Yves von Cramon, and Angela D. Friederici. Revisiting the role of Broca's area in sentence processing: Syntactic integration versus syntactic working memory. *Human Brain Mapping*, 24:79–91, 2005.

[29] P C. Fletcher, F. Happe, U. Frith, S C. Baker, R J. Dolan, R S. Frackowiak, and C D. Frith. Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, 57:109–128, 1995.

[30] Stephani Foraker and Brian McElree. The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, 56:357–383, 2007.

[31] Angela D. Friederici, Michiru Makuuchi, and Jörg Bahlmann. The role of the posterior superior temporal cortex in sentence comprehension. *Neuroreport*, 20:563–568, 2009.

[32] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora*, volume 71, page 76, 1998.

[33] N. Geschwind. Disconnection syndromes in animals and man. *Brain*, 88: 237–294, 585–644, 1965.

[34] Peter. C. Gordon and Randall. Hendrick. The representation and processing of coreference in discourse. *Cognitive Science*, 22:389–424, 1998.

[35] Peter C. Gordon, Barbara J. Grosz, and Laura A. Gilliom. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–47, 1993.

[36] Paul Grice. Logic and conversation. In *Syntax and semantics: Speech acts*.

[37] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *COLING 1996: The 16th International Conference on Computational Linguistics*, volume 1, pages 466–471. Association for Computational Linguistics, 1996.

[38] Yosef Grodzinsky and Tanya Reinhart. The innateness of binding and coreference. *Linguistic Inquiry*, 24:69–101, 1993.

[39] Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21:203–225, 1995.

[40] Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307, 1993.

[41] John Hale. *Automaton theories of human sentence comprehension*. CSLI Publications, 2014.

[42] Anke Hammer, Rainer Goebel, Jens Schwarzbach, Thomas F. Münte, and Bernadette M. Jansma. When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research*, 1146:185–198, 2007.

[43] Anke Hammer, Bernadette M. Jansma, Monique Lamers, and Thomas F. Münte. Interplay of meaning, syntax and working memory during pronoun resolution investigated by ERPs. *Brain Research*, 1230:177–191, 2008.

[44] Anke Hammer, Bernadette M. Jansma, Claus Tempelmann, and Thomas F. Münte. Neural mechanisms of anaphoric reference revealed by fMRI. *Frontiers in Psychology*, 2:1–9, 2011.

[45] Tony Harris, Ken Wexler, and Phillip Holcomb. An ERP investigation of binding and coreference. *Brain & Language*, 75:313–46, 2000.

[46] Stefan Heim. Syntactic gender processing in the human brain: A review and a model. *Brain and Language*, 106:55–64, 2008.

[47] Stephan Heim, B. Opitz, and Angela D. Friederici. Broca's area in the human brain is involved in the selection of grammatical gender for language production: Evidence from event-related functional magnetic resonance imaging. *Neuroscience Letters*, 328:101–104, 2002.

[48] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8:393–402, 2007.

[49] Jerry Hobbs. Resolving pronouns. In *Readings in natural language processing*. Morgan Kaufman Publishers, Inc., Los Altos, California, USA., 1977.

[50] Fumitaka Homae, Noriaki Yahata, and Kuniyoshi Sakai. Selective enhancement of functional connectivity in the left prefrontal cortex during sentence processing. *NeuroImage*, 20:578–586, 2003.

[51] C.-T. James Huang. Pro-drop in Chinese: A generalized control theory. In Osvaldo Jaeggli and Kenneth Safir, editors, *The null subject parameter*, pages 185–214. Springer, 1989.

[52] C.-T. James Huang and C.-Y. Yang Barry. Topic-drop and MCP. In *87th Annual Meeting of the Linguistic Society of America*, 2013.

[53] Scott A. Huettel, Allen W. Song, and Gregory McCarthy. Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices. *Journal of Neuroscience*, 25:3304–3311, 2005.

[54] Colin Humphries, Jeffrey R. Binder, David A. Medler, and Einat. Liebenthal. Time course of semantic processes during sentence comprehension: An fmri study. *NeuroImage*, 36:924–932, 2007.

[55] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458, 2016.

[56] P. Indefrey and Willem Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92:101–144, 2004.

[57] Juhani Järvikivi, Roger P.G. van Gompel, Jukka Hyönä, and Raymond Bertram. Ambiguous pronoun resolution: contrasting the first-mention and subject-preference accounts. *Psychological Science*, 16:260–264, 2007.

[58] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice-Hall, Inc., Pearson Higher Education, New Jersey, 2008.

[59] Varada Kolhatkar and Graeme Hirst. Resolving shell nouns. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, page 499–510. Association for Computational Linguistics, 2014.

[60] Prantik Kundu, Souheil J. Inati, Jennifer W. Evans, Wen-Ming Luh, and Peter A. Bandettini. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage*, 60:1759–1770, 2012.

[61] Gina R. Kuperberg, Balaji M. Lakshmanan, David N. Caplan, and Phillip J. Holcomb. Making sense of discourse: An fMRI study of causal inferencing across sentences. *NeuroImage*, 33:343–361, 2006.

[62] Matthew A. Lambon Ralph, Karen Sage, and Jo Roberts. Classical anomia: A neuropsychological perspective on speech production. *Neuropsychologia*, 38:186–202, 2000.

[63] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20:535–561, 1994.

[64] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics, 2017.

[65] Richard L. Lewis and Shravan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29: 375–419, 2005.

[66] Charles N. Li and Sandra A. Thompson. Subject and topic: A new typology. In Charles N. Li, editor, *Subject and topic*, pages 457–89. Academic Press, New York, USA, 1976.

[67] O. Longe, B. Randall, E. Stamatakis, and L. Tyler. Grammatical categories in the brain: The role of morphological structure. *Cerebral Cortex*, 17: 1812–1820, 2007.

[68] Ana Marasović, Leo Born, Juri Opitz, and Anette Frank. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017*

Conference on Empirical Methods in Natural Language Processing, pages 221–232. Association for Computational Linguistics, 2017.

[69] David Marr. *Vision: A computational approach*. Freeman & Co., San Francisco, 1982.

[70] William Matchin, Jon Sprouse, and Gregory Hickok. A structural distance effect for backward anaphora in Broca's area: An fMRI study. *Brain and Language*, 138:1–11, 2014.

[71] Samuel M. McClure, David I. Laibson, George Loewenstein, and Jonathan D. Cohen. Separate neural systems value immediate and delayed monetary rewards. *Science*, 306:503–507, 2004.

[72] Corey T. McMillan, Robin Clark, Delani Gunawardena, Neville Ryant, and Murray Grossman. fMRI evidence for strategic decision-making during resolution of pronoun reference. *Neuropsychologia*, 50:674–687, 2012.

[73] Gabriele Miceli, Laura Giustolisi, and Alfonso Caramazza. The interaction of lexical and non-lexical processing mechanism: Evidence from anomia. *Cortex*, 27:57–80, 1991.

[74] Gabriele Miceli, Patrizia Turriziani, Carlo Caltagirone, Rita Capasso, Francesco Tomaiuolo, and Alfonso Caramazza. The neural correlates of grammatical gender: An fMRI investigation. *Journal of Cognitive Neuroscience*, 14:618–628, 2002.

[75] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[76] Allen Newell. *Unified theories of cognition.* Harvard University Press, Cambridge, MA, 1990.

[77] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics, 2002.

[78] Mante S. Nieuwland and Jos A. Van Berkum. Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, 1118:155–67, 2006.

[79] Mante S. Nieuwland, Karl M. Petersson, and Jos A. Van Berkum. On sense and reference: Examining the functional neuroanatomy of referential processing. *Neuroimage*, 37:993–1004, 2007.

[80] Lee Osterhout and Linda A. Mobley. Event-related brain potentials elicited by failure to agree. *Journal of Memory & Language*, 34:739–73, 1995.

[81] Lee Osterhout, Michael Bersick, and Judith McLaughlin. Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, 25:273–85, 1997.

[82] Jonathan W. Peirce. PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods*, 162:8–13, 2007.

[83] William Penny, Karl Friston, John Ashburner, Stefan Kiebel, and Thomas Nichols. *Statistical parametric mapping: The analysis of functional brain images*. Academic Press., 2011.

[84] Peter L. Pirolli and John R. Anderson. The role of practice in fact retrieval.

Journal of Experimental Psychology: Learning, Memory and Cognition, 11:126–153, 1985.

[85] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning*, pages 1–40. Association for Computational Linguistics, 2012.

[86] Amy R. Price, Michael F. Bonner, Jonathan E. Peelle, and Murray Grossman. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *The Journal of Neuroscience*, 35:3276–3284, 2015.

[87] Tanya Reinhart. *Anaphora and semantic interpretation*. Croom Helm, London, 1983.

[88] Tanya Reinhart and Eric Reuland. Reflexivity. *Linguistic Inquiry*, 24:657–720, 1993.

[89] Eric Reuland. Primitives of binding. *Linguistic Inquiry*, 32:439–492, 2001.

[90] Luigi Rizzi. Null objects in Italian and the theory of pro. *Linguistic Inquiry*, 17:501–557, 1986.

[91] Anthony J. Sanford, K. Moar, and Simon C. Garrod. Proper names as controllers of discourse focus. *Language and speech*, 31:43–56, 1988.

[92] Andrea Santi and Yosef Grodzinsky. Working memory and syntax interact in Broca's area. *NeuroImage*, 37:8–17, 2007.

[93] Andrea Santi and Yosef Grodzinsky. Broca's area and sentence comprehension: A relationship parasitic on dependency, displacement or predictability? *Neuropsychologia*, 50:821–832, 2012.

[94] Bernadette M. Schmitt, Monique Lamers, and Thomas F. Münte. Electrophysiological estimates of biological and syntactic gender violation during pronoun processing. *Cognitive Brain Research*, 14:333–46, 2002.

[95] Sophie K. Scott, C. Catrin. Blank, Stuart. Rosen, and Richard J. S. Wise. Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123:2400–2406, 2000.

[96] Myeong-Ho Sohn, Adam Goode, Andrew V. Stenger, Cameron S. Carter, and John R. Anderson. Competition and representation during memory retrieval: Roles of the prefrontal cortex and the posterior parietal cortex. *Proceedings of the National Academy of Sciences of the USA*, 100:7412–7417, 2003.

[97] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27:521–544, 2001.

[98] Galina Spitsyna, Jane E. Warren, Sophie K. Scott, Federico E. Turkheimer, and Richard J. S. Wise. Converging language streams in the human temporal lobe. *Journal of Neuroscience*, 26:7328–7336, 2006.

[99] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of*

the European Chapter of the Association for Computational Linguistics, pages 102–107. Association for Computational Linguistics, 2012.

[100] Tamara Y. Swaab, C. Christine Camblin, and Peter C. Gordon. Electrophysiological evidence for reversed lexical repetition effects in language processing. *Journal of Cognitive Neuroscience*, 16:715–726, 2004.

[101] Tarald Taraldsen. On the NIC, vacuous application and the that-trace filter. 1978.

[102] Fengfu Tsao. *A functional study of topic in Chinese: The first step towards discourse analysis*. PhD thesis, USC, Los Angeles, California, 1977.

[103] Lorraine K. Tyler, William D. Marslen-Wilson, Billi Randall, Paul Wright, Barry J. Devereux, Jie Zhuang, Papoutsi Marina, and Emmanuel A. Stamatakis. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain*, 134:415–431, 2011.

[104] Jos J. A. van Berkum, Colin M. Brown, and Peter Hagoort. Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory & Language*, 41:147–82, 1999.

[105] Jos J. A. van Berkum, Colin M. Brown, Peter Hagoort, and Pienie Zwitserlood. Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. *Psychophysiology*, 40:235–48, 2003.

[106] Jos J. A. van Berkum, Arnout W. Koornneef, Marte Otten, and Mante S. Nieuwland. Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, 1146:158–71, 2007.

[107] Jacolien van Rij, Hedderik van Rijn, and Petra Hendriks. How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, 5:564–580, 2013.

[108] Timothy J. Vickery and Yuhong V. Jiang. Inferior parietal lobule supports decision making under uncertainty in humans. *Cerebral Cortex*, 19:916–925, 2009.

[109] M. Visser and M. Lambon Ralph. Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of Cognitive Neuroscience*, 23:3121–3131, 2011.

[110] Qi Wang, Diane Lillo-Martin, Catherine T. Best, and Andrea Levitt. Null subject versus null object: Some evidence from the acquisition of Chinese and English. *Language Acquisition*, 2:221–254, 1992.

[111] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9: e112575, 2014.

[112] Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. Learning global features for coreference resolution. pages 994–1004, 2016.

[113] Jiang Xu, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, 25:1002–1015, 2005.

[114] Satoru Yokoyama, Hideyuki Okamoto, Tadao Miyamoto, Kei Yoshimoto, Jungho Kim, Kazuki Iwata, Hyeonjeong Jeong, Shinya Uchida, Naho Ikuta,

Yuko Sassa, Wataru Nakamura, Kaoru Horie, Shigeru Sato, and Ryuta Kawashimab. Cortical activation in the processing of passive sentences in l1 and l2: An fmri study. *NeuroImage*, 30:570–579, 2006.